HIGH SCHOOL STATISTICS TEACHERS' UNDERSTANDING OF HYPOTHESIS TESTING THROUGH SIMULATION

by

Amber L. Matuszewski

A Dissertation Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy in Mathematics and Science Education

Middle Tennessee State University
August 2018

Dissertation Committee:

Dr. Jeremy Strayer, Chair

Dr. Sarah Bleiler-Baxter

Dr. Seth Jones

Dr. Jennifer Lovett

Dr. Ginger Rowell

I dedicate this dissertation to my husband, Steve Matuszewski. Thank you for encouraging me and always believing in me. I would not have accomplished half of what I have without you.

ACKNOWLEDGEMENTS

First, I would like to thank my amazing family for being there for me these last four years. Teaching full time while completing this program would not have been possible without you. Thank you to my husband for cleaning the house, going grocery shopping, and taking care of pretty much everything else. Especially me, when I was having a particularly tough time. Thank you to my parents, Brown and Tina Sanford, for your constant love and support, along with the Sunday dinners when I needed them. Thank you to my son, Clifton, who also helped take care of the house and who never failed to tell me how proud he was of me.

I also want to thank my chair, Dr. Jeremy Strayer, for introducing me to the world of simulations for inference. Thank you for always pushing me, but also for always believing in me. Without your knowledge, feedback, and encouragement, this dissertation would not have been possible. Also, thank you to my committee members, Dr. Sarah Bleiler-Baxter, Dr. Seth Jones, Dr. Jennifer Lovett, and Dr. Ginger Rowell. Your insights and feedback were invaluable to this project. I truly appreciate you all and could not have selected a better committee.

Finally, thank you to my writing group, Dr. Matt Duncan and Dr. Candace Terry. We did it. Also, thank you to Lucy Watson. You are next.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

It is a growing trend in statistics education to use simulations for hypothesis testing due to the belief that simulations help make the abstract ideas behind hypothesis testing become more concrete and understandable. Using simulations involves randomizing data production, repeating by simulation to see what is typical, and rejecting the null hypothesis if your data falls in the tails of the simulated distribution. This explanatory multiple case study sought to answer the following questions, "How does engaging in simulation tasks for hypothesis testing influence high school statistics teachers' understanding of traditional hypothesis testing?" and "How do simulation tasks influence high school statistics teachers' understanding of simulations and how do they make connections between traditional and simulation approaches for hypothesis testing?" The results of this study revealed that teachers' understanding of hypothesis testing was positively impacted because of engaging in simulation tasks. The focus of the simulation tasks on concepts and logic instead of procedures helped the participants develop their understanding of the logic of hypothesis testing. Additionally, simulation approaches focus on visualizations, which helped develop the understanding of the probabilistic nature of hypothesis testing. Finally, how teachers understood simulations and made connections between traditional and simulation approaches was directly influenced by the lesson plan design.

CHAPTER ONE: INTRODUCTION

Introduction

People often make decisions based on statistical information communicated on television, in the newspaper, or online (Franklin et al., 2007). The Statistics Education of Teachers (SET) (Franklin et al., 2015) document stated,

> In an increasingly data-driven world, statistical literacy is becoming an essential competency, not only for researchers conducting formal statistical analyses, but for informed citizens making everyday decisions based on data. Whether following media coverage of current events, making financial decisions, or assessing health risks, the ability to process statistical information is critical for navigating modern society. (p.1)

Due to the importance of developing statistically literate members of society, the last several decades have seen a drive by reputable organizations, such as the National Council of Teachers of Mathematics (NCTM) and the American Statistical Association (ASA), to explicate the importance of the inclusion of statistical topics as early as elementary school and building on these foundational ideas throughout students' educational experiences. NCTM began this push with their release of *Curriculum and Evaluation Standards for School Mathematics* in 1989 and further emphasized the importance of statistics in the 2000 publication of NCTM's *Principles and Standards for School Mathematics,* with statistics as one of the five major content strands. Following this publication, the Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report (Franklin et al., 2007), endorsed by the American Statistical Association (ASA), was released with the intention of complementing NCTM's recommendations

(Franklin & Kader, 2010) and offered a framework for teaching statistics. The commonality among all these documents is a focus on including more statistical topics in the curriculum.

With recommendations to include more statistical topics throughout students' educational experiences, there has also been an increase of college majors requiring at least one course in statistics. This has led to a dramatic increase in the enrollment of students taking Advanced Placement (AP) Statistics at the high school level (Franklin et al., 2011), and more high schools are also offering a similar non-AP stand-alone statistics course. The AP Statistics course is like an introductory-level college statistics course and addresses four broad conceptual themes, which include exploring data, sampling and experimentation, anticipating patterns, and statistical inference (College Board, 2010). The AP exam consists of a mix of multiple choice and free response questions, with a focus on conceptual understanding (Franklin et al., 2011). Scores range from one to five, and students will typically receive college credit for a score of three or higher, although individual universities set their own acceptance scores (College Board, 2010).

The first AP Statistics exam was given to approximately 7,500 students and has now grown to over 100,000 students taking the exam each year. With this increase in enrollment, more AP Statistics teachers are needed. Typically, AP Statistics teachers are secondary mathematics teachers, of which there is a general concern that these teachers are underprepared and have few experiences with statistics themselves (Ben-Zvi & Garfield, 2004). Teachers of non-AP stand-alone statistics courses are usually high school mathematics teachers as well. Statistics is taught by mathematics teachers, because there is a general understanding that statistics is part of the mathematics

curriculum. However, statisticians vehemently defend the idea of statistics being its own unique discipline (Scheaffer, 2006).

Statistics, like many other disciplines, uses mathematics as a tool, but there are many things that distinguish statistics from mathematics. For example, statistical thinking and mathematical thinking are quite different (Wild & Pfannkuch, 1999). Mathematical thinking involves focusing on logic and proof. However, statistical thinking includes the recognition of the need for data, the consideration of looking at data from different perspectives, an attention to variation, contemplation of statistical models, and a focus on integrating the statistical and contextual (Wild & Pfannkuch, 1999). In fact, the role of context is seen as critical in statistics (Cobb & Moore, 1997). Statistics is about data in context, while mathematics focuses on abstraction and proof (Scheaffer, 2006).

In addition to kinds of thinking and importance of context, the type of answer sought in each field is also different. Generally, in a mathematics classroom, you are trying to determine the final correct solution and are able to check the answer. However, in statistics, you are trying to arrive at an answer in which you can feel a certain amount of confidence. Statistics uses probabilities to make decisions and recognizes that the answer may not be correct. Scheaffer (2006) described this as deterministic thinking in mathematics and probabilistic thinking in statistics. Even the problem-solving process is different between the two disciplines. A typical mathematical problem-solving approach is exemplified by Polya's (1945) process of understand the problem, devise a plan, carry out the plan, and look back at your results. In contrast, Franklin et al. (2007) listed the four components of statistical problem solving as formulate questions, collect data, analyze data, and interpret results. Additionally, there is a focus on variability under each

component. Under the formulating questions section, one is said to anticipate variability. While collecting data, variability is to be acknowledged. Under analyzing data, variability is accounted for; and, finally, when interpreting results, one must allow for variability. As discussed in this section, although statistics has been taught through mathematics departments by tradition, the field of statistics is quite different from mathematics.

These differences are important, and both AP and non-AP Statistics teachers should be prepared to deal with these paradigm shifts between their mathematical and statistical content areas. However, mathematics teachers are offered little to no help in understanding these differences. Additionally, most teachers lack a fundamental understanding of the basics in this subject (Franklin, 2013). Makar and Confrey (2004) found that in-service teachers struggled with the concept of sampling distributions, even after a six-month professional development project. Lovett and Lee (2017) found that pre-service mathematics teachers (PSMT) showed weaknesses in understanding variability, sampling distributions, *p*-values, and confidence intervals. Additionally, these PSMT did not feel confident in teaching these topics. These topics comprise the foundation of understanding inference. Inferential statistics refers to drawing or making conclusions from data and includes both confidence intervals and hypothesis testing (Starnes, Yates, & Moore, 2010), which the AP curriculum dedicates 30-40% of its content (College Board, 2010). Not only does AP Statistics include these challenging topics, but the AP Statistics curriculum designers incorporated statistical topics that are often not encountered until a second course in statistics in college (Franklin et al., 2011). The depth of knowledge required to teach AP Statistics goes beyond what secondary mathematics teachers typically acquire through their coursework. Additionally, the AP

exam focuses on testing students' conceptual understanding of these topics, which many of the AP Statistics teachers may lack due to insufficient training (Franklin et al., 2011). In fact, any high school teacher of a stand-alone statistics course would also need additional training to teach towards understanding versus simply memorizing facts and procedures (Franklin et al., 2015).

Therefore, due to the general acknowledgement that these high school mathematics teachers who are teaching these statistic courses lack the necessary training, it is essential to help them foster a deep understanding of these inferential topics, along with equipping them with effective techniques that will help them develop a deep understanding in their own students. One area of importance in inferential statistics is that of hypothesis testing, which is a major component of introductory statistics courses. This topic has been shown to be one of the most difficult topics for students and teachers alike (Smith, 2008; Thompson, Liu, & Saldanha, 2007).

A traditional hypothesis test is used to evaluate a claim. By determining two competing hypotheses, indirect logic is employed to test the claim. The claim that one is trying to gather evidence for through a random sample or experiment is called the alternative hypothesis. The competing hypothesis, which is typically a statement of no effect or no difference, is called the null hypothesis. One assumes that the null hypothesis is true, and by looking at what the typical results would be under that assumption, one can determine if the results from the study or experiment are surprising enough to be able to reject the null hypothesis. A *p*-value is used to assess how surprising the results are. A *p*-value is the probability of obtaining results as extreme or more extreme as the ones you obtained, if the null hypothesis is true. If the results are very surprising given that the null

hypothesis is true, meaning your data has a small probability of occurring, then this is evidence that the null hypothesis is not true. To determine if the results are surprising enough, researchers typically compare the *p*-value to an alpha-level, which is also called the significance level. Commonly, an alpha-level of .05 is used, although the context of one's study should influence the choice of the significance level. If the *p*-value is less than alpha, then one would reject the null hypothesis. If the conclusion is to reject the null hypothesis, then the logic states that there is convincing evidence to conclude your alternative hypothesis.

For example, a possible hypothesis may be that the majority of students at your school favor the current president. For this scenario, the null hypothesis would be that the proportion of students at your school who favor the president is 50%. The alternative would be that the proportion of students who favor the president at your school is greater than 50%. One would then obtain a sample of students from your school and determine the percent that favor the president. With a traditional test, one would calculate the standardized sample proportion, called a *z*-test statistic, and determine the probability of obtaining that statistic or higher if the true proportion is 50%. Although the logic behind the test is simple, traditional approaches involve complicated formulas and the use of theoretical sampling distributions. With the general acknowledgement that many students simply memorize steps, there is a call for students to develop a deeper understanding of these topics by learning how to reason and think statistically (Ben-Zvi & Garfield, 2004).

However, only recently has research emerged regarding the development of statistical reasoning and thinking (e.g. Reading & Shaughnessy, 2004; Moritz, 2004; Watson, 2004). To help students develop these abilities, they need opportunities to

visualize and explore. Therefore, teachers should use appropriate technology and active learning approaches (Garfield & Ben-Zvi, 2007). For hypothesis testing, using simulations and technology to teach inference allows these ideas to become more concrete and understandable to students (Erickson, 2006). Lane-Getaz (2010) found that students who learned these concepts through simulation activities statistically out-performed students with less simulation exposure regarding inferential statistical topics. In fact, Cobb (2007) argued for introductory statistics courses in college to focus solely on using the logic behind inference and simulations to teach this topic versus the traditional method of using theoretical sampling distributions and formulas. He stated that the logic behind a hypothesis test was based on "the three R's: randomize, repeat, reject. Randomize data production; repeat by simulation to see what's typical; reject any model that puts your data in its tail" (Cobb, 2007, p. 12). This equates to using simulated sampling distributions to reject or fail to reject the null hypothesis, which I will refer to as a simulation approach for hypothesis testing. The main difference between a traditional approach and a simulation approach is that traditional approaches use formulas to calculate a test statistic and obtain a *p*-value based on a theoretical sampling distribution. Additionally, conditions must be met to ensure that your calculations will be accurate. For example, if your population is known to be non-normal, then your sample size needs to be large enough so that the theoretical sampling distribution is approximately normal. However, simulation approaches for hypothesis testing do not require checking assumptions. Rather, they involve creating an empirical sampling distribution using simulation, often with technology, and determining if your sample data is surprising, which means that it falls in the tails of the distribution. To illustrate this concept, I have

provided Lane-Getaz and Zieffler's Simulation Process Model (SPM) (2006) for this approach (*Figure 1*), and I will reference this model as I describe how to work the hypothesis example described above with a simulation approach.

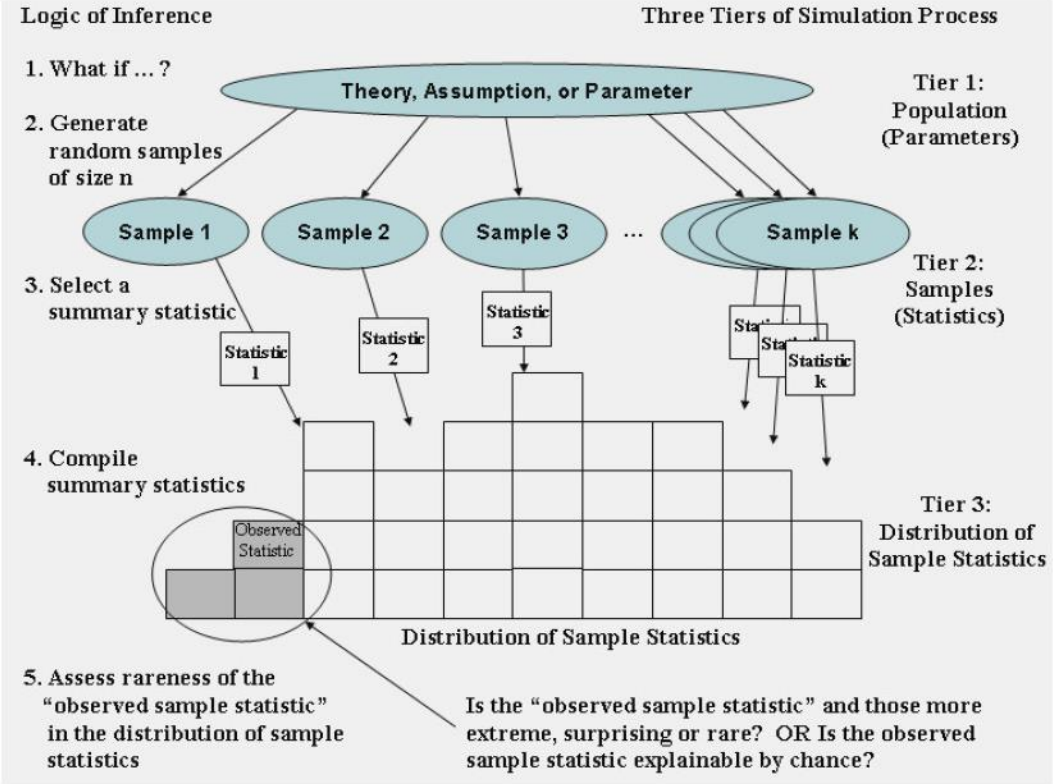


Figure 1. Simulation Process Model (Lane & Zieffler, 2006)

The simulation approach would be used instead of using formulas and theory to calculate a *p*-value, as described under the traditional approach. The left-hand side of the SPM illustrates five steps behind the logic of a hypothesis test, and the right-hand side describes the three tiers of using a simulation process. In the previous hypothesis test example, the logic behind this approach asks what would happen if 50% of the students at

a school did approve of the president. Tier one of the simulation approach would define the null hypothesis as $p = .5$, where p = the proportion of students at the school who are in favor of the president. As an example, for this scenario, I will assume that I sampled 20 students and obtained a sample proportion of 75%. This proportion seems surprising if only 50% approve the current president, but a simulation in which I assume that the null hypothesis is true would allow me to see what types of sample proportions are typical under that model and determine if my sample proportion is surprising or not based on probability. This leads to steps two and three of the logic steps, which would be to generate samples of size 20, assuming $p = .5$, and calculating many sample proportions. Tier two of the simulation process involves the production of these samples. Because the null hypothesis is $p = .5$, I could use a coin to simulate that process. Heads could represent a student being in favor of the president, and tails would represent not being in favor. Additionally, because my sample size was 20, I would need to flip the coin 20 times to represent one trial and determine the proportion of times the coin landed on heads. To see what types of results are typical and obtain a more accurate estimation of the *p*-value, many trials would need to be conducted. I could use technology to simulate the trials multiple times and produce a simulated sampling distribution of sample proportions, which corresponds to the logic step four and to tier three of the simulation process. The final step of a simulation approach would be to obtain an approximate *p*-value by seeing how many times the simulated samples had a sample proportion of 75% or higher. This would allow me to determine if my result was surprising enough to reject the null hypothesis and if I have enough evidence to conclude the alternative. This approach simplifies the process and allows the logic of a hypothesis test to take

precedence over procedures. One no longer must use formulas to calculate the test statistic or check conditions to use a theoretical sampling distribution. Using these techniques may allow teachers to not only help their own students comprehend the logic behind a hypothesis test, but to also develop their own understanding of hypothesis testing.

Problem Statement

Despite the difficulty in learning and teaching the topic of hypothesis testing, few studies exist exploring high school teachers' understanding of this topic (e.g. Peters, 2009). Additionally, very few studies have investigated students' understandings of simulation approaches (e.g. Lane-Getaz, 2010; Saldanha, & Thompson, 2014), and no studies have investigated high school statistics teachers' understanding of traditional hypothesis testing through simulation. However, if high school statistics teachers begin to use this technique in their classroom to foster understanding, it is important to know how they understand these simulation approaches and how they connect their current knowledge of traditional approaches with this understanding.

Purpose of the Study

The purpose of this study was to investigate how high school statistics teachers' understanding concerning hypothesis testing was influenced when these teachers engaged in simulation tasks for hypothesis testing. Additionally, I wished to gain insight into how these teachers understood simulations and how they connected traditional and simulation approaches. My population of interest for this study included both high school AP Statistics teachers and other high school mathematics teachers of stand-alone statistics classes with a similar curriculum, which includes statistical inference. For simplicity, I

will refer to these teachers as high school statistics teachers for the remainder of this dissertation. I used a multiple-case study design and selected high school statistics teachers who had not previously used simulation approaches, as my cases. I sought to answer the following questions:

1. How does engaging in simulation tasks for hypothesis testing influence high school statistics teachers' understanding of traditional hypothesis testing?
2. How do simulation tasks influence high school statistics teachers' understanding of simulations and how do they how do they make connections between traditional and simulation approaches for hypothesis testing?

Significance

By using simulations to introduce the concept of a hypothesis test, students can gain a greater conceptual understanding of this topic (Lane-Getaz, 2010). With the consensus that most mathematics teachers, which includes high school statistics teachers, lack the conceptual understanding of many foundational topics in statistics, it is critical to learn how to develop this type of understanding not only in students, but in teachers as well. This study sought to investigate how using simulations can foster this understanding in high school statistics teachers. This is important because possessing strong content knowledge is an important component of the mathematical knowledge needed for teaching (Ball, Thames, & Phelps, 2008). Understanding how teachers think and how to foster a deeper understanding of hypothesis testing, a fundamental statistical topic, can help guide professional development for teachers in the future. Additionally, as

recommended by Lane-Getaz and Zieffler (2006), more studies are needed to see how using simulations for inference can help develop a deeper conceptual understanding of traditional hypothesis testing.

Definitions

To support the clarity of this dissertation, the following definitions are offered.

Approximate Sampling Distribution

In contrast to a theoretical sampling distribution, which is comprised of all possible samples of the same size from a given population, an approximate sampling distribution is constructed by simulating the sampling process. Both distributions show how a statistic varies from sample to sample (Peck, Gould, & Miller, 2013).

Common Content Knowledge (CCK)

Common content knowledge is the basic knowledge of the subject matter that is not specific to teaching. It involves being able to correctly work problems and use appropriate terms and notation (Ball et al., 2008).

High school statistics teacher

For this study, high school statistics teacher will refer to any secondary mathematics teacher who teaches AP Statistics or a similar non-AP statistics class, which includes inferential statistics topics.

Key Developmental Understandings (KDUs)

KDUs are used to identify conceptual learning goals in mathematics. This is not knowledge that can simply be explained or demonstrated to the student. Instead, KDUs represent a conceptual advancement which must be developed over time and which allow the student to perceive the underlying mathematical relationships. This type of

understanding cannot be acquired through explanation or demonstration. Instead, multiple exposures to activities and reflections must be used. Students must be observed in action of the task to identify the corresponding KDUs (Simon, 2006).

Simulation Approach for Hypothesis Testing

In contrast to traditional approaches, this technique involves using simulation to produce repeated samples from a model representing the null hypothesis and to reject the null hypothesis if sample data falls in the tails of the simulated sampling distribution (Cobb, 2007). For this study, a simulation approach is used as a pedagogical tool to promote a deeper understanding of traditional hypothesis testing.

Specialized Content Knowledge (SCK)

The content knowledge that is unique to teaching is referred to as specialized content knowledge. This knowledge is typically only needed for teachers. It goes beyond a basic understanding of how to correctly work a problem. A teacher must also be able to determine if different solution paths will work and understand why they work. (Ball et al., 2008).

Statistical Inference

Inference is the process of using data from a sample or experiment to make a decision or draw a conclusion (Peck et al., 2013).

Traditional Approach for Hypothesis Testing

A traditional approach to hypothesis testing will typically follow a similar pattern to the steps listed: 1) Determine appropriate null and alternative hypothesis 2) Determine appropriate procedure, such as a one sample *t*-test for means, and check that conditions are met for the determined procedure 3) Calculate test statistic and *p*-value 4) Draw a

conclusion based on the *p*-value (Starnes et al., 2010). This method to hypothesis testing which is taught in most classrooms combines the Neyman-Person and Fisher approaches and is identified by statisticians as null hypothesis significance testing.

Chapter Summary

In conclusion, simulation techniques are gaining popularity due to advances in technology and the importance of developing a deeper understanding of statistical topics in contrast to simply memorizing steps and procedures. However, little is known about how using these techniques may help foster high school statistics teachers' understanding of hypothesis testing. My dissertation sought to gain insight into how simulation approaches nurture this development and how high school statistics teachers, who have not previously learned about simulations for inference, understand simulations and make connections between the traditional and simulation techniques.

CHAPTER TWO: REVIEW OF LITERATURE

Introduction

The purpose of this study was to investigate how high school statistics teachers' understanding of hypothesis testing was influenced when these teachers engaged simulation tasks for hypothesis testing. Additionally, I sought to gain insight into how these teachers understood simulations and connected traditional and simulation approaches. My specific research questions are:

1. How does engaging in simulation tasks for hypothesis testing influence high school statistics teachers' understanding of traditional hypothesis testing?
2. How do simulation tasks influence high school statistics teachers' understanding of simulations and how do they make connections between traditional and simulation approaches for hypothesis testing?

To answer my research questions, it is essential to review what the literature states about understanding hypothesis testing. Therefore, in the first section of Chapter Two, I will discuss the literature related to understanding hypothesis testing and its associated concepts. Next, I will discuss the relevant research concerning teachers' understanding of hypothesis testing. I will continue with a discussion of simulations for inference. Finally, I will conclude by describing the theoretical framework that was used to guide this study.

Hypothesis Testing and Related Concepts

Although this study focused on teachers' understanding of hypothesis testing, literature concerning teachers' understanding of this topic is limited. In this section, I will use studies related to students' understandings to discuss the specific related concepts of

hypothesis testing. Additionally, including studies related to students is appropriate because teachers often possess an understanding of hypothesis testing similar to students (Harradine, Batanero, & Rossman, 2011).

The organization of this section is based on the big ideas and essential understandings related to hypothesis testing discussed in *Developing Essential Understandings of Statistics* (Peck et al., 2013). Big idea number three concerned hypothesis testing and stated, "Hypothesis tests answer the question do I think this result could have happened by chance" (Peck et al., 2013, p. 44). This idea concerned the overall logic of hypothesis testing. Big idea two concerned variability, including sampling distributions, and big idea four concerned data collection. Although probability is not listed as a big idea, the importance of *p*-value and significance level were discussed. Therefore, based on these big ideas and essential understandings, I have divided this section about hypothesis testing into the categories of the logic of hypothesis testing, probability, data collection, variability, and sampling distribution. I will discuss the first three of these topics individually and then combine the topics of variability and sampling distribution, because they are difficult to discuss separately regarding hypothesis testing. To facilitate the flow of this section and show how each concept is related to hypothesis testing, I will reference the following hypothesis test example.

Hypothesis Test Example

I will use the hypothesis test example that corresponds to the third simulation task used in this study. The example references an experiment conducted by Antonioli and Reveley (2005), which investigated the effectiveness of dolphin therapy to relieve mild to moderate depression. Thirty participants were randomly assigned to one of two groups.

The fifteen subjects in the control group swam and snorkeled each day. The other fifteen participants in the treatment group also swam and snorkeled each day, but they did so in the presence of bottlenose dolphins. At the end of the study, the subjects' level of depression was reevaluated to determine if they showed substantial improvement. For the control group, 3 out of 15 showed substantial improvement, in comparison to 10 out of 15 in the treatment group. A two-sample *z*-test for proportions could be used to see if the results are statistically significant. I will discuss this example in terms of each hypothesis test concept in the following sections.

Logic of Hypothesis Testing

Hypothesis testing is based on indirect logic, which is similar to proof by contradiction (Thompson et al., 2007). In mathematics, proof by contradiction involves assuming the negation of the proposition that you are trying to prove is true. However, if assuming this leads to a contradiction, then you have now shown that your original proposition is true. In statistics, hypothesis testing is not a means of proof, but of establishing if the data provide convincing evidence or not for some claim (Peck et al., 2013). Still based on indirect logic, hypothesis testing dictates the need for two competing hypotheses, called the null and alternative, instead of propositions. Typically, what the researcher is trying to obtain evidence for is called the alternative hypothesis, denoted H_a. The alternative hypothesis for the dolphin example would be that the proportion of improvers for the dolphin treatment is greater than the proportion of improvers for the control treatment. Notice the direction of the alternative, in this case greater than, is determined by the practical needs of the researcher. The null hypothesis is typically stated in terms of no change or no difference and is denoted H_o. For the dolphin

example, the null hypothesis would be that the treatments would result in the same proportion of participants showing substantial improvement.

After establishing the null and alternative hypothesis, indirect reasoning asks the question, “Assuming the null hypothesis is true, could my results have happened by chance?” However, a hypothesis test does not determine the likelihood of the null hypothesis being true. The test determines the likelihood of obtaining results as extreme or more extreme than the one we got if the null hypothesis is true. Like reaching a contradiction in mathematics, statistical significance is achieved if the data were inconsistent with results expected under the null hypothesis being true, due to chance variation. However, this type of logic is often not employed by people in everyday life. Studies have shown that people tend to look for confirming evidence, rather than disconfirming evidence, which is referred to in psychology as confirmation bias. (Nickerson, 1998). In fact, “confirmation bias is perhaps the best known and most widely accepted notion of inferential error to come out of the literature on human reasoning” (Evans, 1989, p. 41).

This type of bias has its roots in cognitive psychology and has been well documented using Wason’s selection task. In this task, participants are shown four cards with numbers (even or odd) on one side and letters (vowel or consonant) on the other side. Participants are then asked to test the veracity of the following statement, “If a card has a vowel on one side then it has an even number on the other side”. Both a card with a vowel and an odd number should be flipped, because they hold the potential to falsify the statement. However, most people are unable to correctly identify the two cards which should be flipped to logically test the claim versus selecting cards which would only

confirm the rule. Seeking to confirm a rule is in direct opposition to the logic of hypothesis testing which seeks evidence of one hypothesis by discrediting a competing hypothesis using probability.

Additionally, students believe that a hypothesis is a form of proof in which you can arrive at a deterministic answer of the null hypothesis being true or false (Batanero, 2000). DelMas, Garfield, Ooms, and Chance (2007) conducted a study using the Comprehensive Assessment of Outcomes in a First Statistics course (CAOS) test, as a pre-and post-test for students taking an introductory-level statistics course at the college level. The test was administered to 1470 college students across 33 universities. Although the number of students who were able to correctly reject the null hypothesis increased statistically, the percent of students who believed that rejecting the null hypothesis meant it was false increased.

Sotos, Vanhoof, Van den Noortgate, and Onghena (2009) also reported this misconception in students who had taken an introductory college-level statistics course. They administered a five-question multiple choice test inspired by items from the ARTIST website, which is the same group that developed the CAOS test used in the study just described. The researchers designed the questions to address misconceptions regarding the concept of a hypothesis test, *p*-values, and significance levels. One hundred and forty-four students at a Spanish university completed the exam at the end of an introductory statistics course. Regarding the logic of hypothesis testing as a means of assessing evidence to reject or fail to reject the null hypothesis, just under half of the students correctly identified this as the only correct purpose of a hypothesis test. Approximately 20% selected the response which indicated that a hypothesis test was a

form of proof, and another 19% selected the response that a hypothesis test determined the probability or improbability of the null hypothesis.

In conclusion, the logic of a hypothesis test is not something that everyday people tend to employ. Even after taking an introductory college-level statistics course, students often fail to correctly identify a hypothesis test as a means of assessing evidence to reject or fail to reject the null hypothesis. Students often believe that a hypothesis test can be used to definitively prove or disprove a claim or that it can be used to determine the probability of the null hypothesis being true or false.

Probability

The second important component of a hypothesis test concerns probability. The role of probability is essential in interpreting the *p*-value and significance level. For the dolphin example, the difference in proportions was approximately .47. The probability of obtaining a difference of .47 or higher, if there is truly no difference in proportions, is theoretically .00495, which is the *p*-value one would obtain from a two-sample *z*-test. Assuming a significance level even as low as .01 would result in rejecting the null hypothesis and having convincing evidence that swimming with dolphins does help reduce depression. The significance level of .01 means that there is a 1 percent chance in this scenario of incorrectly rejecting the null hypothesis when it is true, which is called a Type I error. The significance level directly corresponds to the rejection region and allows the researcher to have control over the probability of committing a Type I error.

Some misconceptions that may arise due to this lack of understanding the role of probability include believing that if you reject the null hypothesis you have proven it wrong (Liu & Thompson, 2009; Krauss & Wassner, 2002) and confusing the *p*-value and

significance level (Lane-Getaz, 2010). Additionally, the *p*-value is often incorrectly interpreted as the probability of the null hypothesis being false (Lane-Getaz, 2007). The correct interpretation of the *p*-value is the conditional probability of obtaining results as extreme or more extreme than the one obtained under the assumption that the null hypothesis is true (Smith, 2008). Also, the significance level is the probability of mistakenly rejecting the null hypothesis when it is true (Smith, 2008).

As stated, holding probability misconceptions can impact the ability to correctly interpret the *p*-value in a hypothesis and can impede the ability to make correct conclusions. Therefore, eliminating probability misconceptions is essential; however, studies have shown that learning probability concepts is often much more challenging than assumed (Garfield & Ben-Zvi, 2007).

For example, in a study with college-age students, Hirsch and O'Donnell (2001) found that probability misconceptions are difficult to eliminate and resistant to typical classroom instruction. After their initial findings, Hirsch and O'Donnell (2001) created instructional interventions to address these problems. The intervention included the use of activities which had students make predictions about probabilities involving drawing marbles and winning at games of chance. This is like the tasks used in this study, in which teachers were asked to make a guess regarding the types of results expected if the null hypothesis were true. Next, students would draw the marbles themselves and simulate the game of chance. Again, the tasks in this study were similar in that teachers would then use simulation to see what types of results would be typical. After seeing the actual results, students discussed if the results were consistent with their predictions. These activities resulted in a cognitive conflict when the results did not match their

predictions. Although initial testing did not reach statistical significance, in a follow up test, Hirsch and O'Donnell (2001) found that the college students engaged in simulation activities statistically outperformed the control group in overcoming misconceptions. This indicated that using simulations and creating cognitive conflict may have more lasting effects at eliminating misconceptions.

Similarly, Garfield and Ahlgren (1988) stated that students were unable to learn probability as it is typically taught and recommended that teachers use activities, simulations, and visual illustrations. Using small groups with activity-based lessons, Shaughnessy (1977) found that college-level students' understanding of probability improved. In these small groups, students explored and discovered probability models and rules on their own. Additionally, students were engaged in activities which had them develop probability models and discover probability rules. The students in the activity-based course outperformed students in a traditional lecture-based course in probability.

Lane and Tang (2000) conducted a study in which college students were taught statistical concepts using a traditional textbook and another group of students were taught these concepts using simulations. The group using simulations performed significantly better in responding to questions involving interpreting and using probabilities. The purpose of the simulations used in the tasks in this study was also to help visualize the *p*-value as a probability and aid in interpreting this value. Additionally, the possibility of committing a Type I error can be seen in the simulated sampling distribution and help overcome the misconception that rejecting the null hypothesis means that it must be false.

Employing an experimental matched-pairs design with seventh grade students, Gurbuz and Birgin (2012) found that computer-assisted teaching was more effective than

traditional methods at correcting students' misconceptions of probability. Computer-assisted teaching referred to classes in which technology was incorporated in the presentation of the material. For example, simulations and animations were used to illustrate theoretical probabilities to help students overcome misconceptions. Just like the previously mentioned studies, the use of simulations has been shown to be more effective than traditional teaching methods, when teaching probability.

In summary, to help students with challenging probability concepts, students should be given the opportunity to confront their misconceptions and construct their own knowledge through activities (Hirsch & O'Donnell, 2001). Additionally, using simulations for inference allows students to see how the sampling distribution is constructed and provides a way to make the abstract concepts of *p*-value and significance level more concrete (Erickson, 2006).

Data Collection

The way data is collected for a hypothesis test influences the type of conclusions that can be drawn (Peck et al., 2013). To quantify uncertainty and determine the *p*-value of the test, knowledge about the sampling distribution must be used. This knowledge is based on some type of chance process being employed. Data can either be collected through random sampling or random assignment. Random sampling allows one to generalize to a population of interest, and random assignment permits causal conclusions (see Table 1). For the dolphin example, participants were not randomly selected. Therefore, one cannot generalize the study's results to all people with mild to moderate depression. However, the study did randomly assign participants to treatments. Thus, one

can conclude that swimming with dolphins causes a reduction in depression in people similar to ones in the study.

Table 1

Conclusions from Data Collection (Peck et al., 2013, p. 66)

	Random Sampling	No Random Sampling
Random Assignment	Can infer causality and can generalize from sample at hand to larger population	Can infer causality, but cannot generalize from sample at hand to larger population
No Random Assignment	Can generalize from sample at hand to larger population, but cannot infer causality	Cannot generalize from sample at hand to larger population and cannot infer causality

Watson and Moritz (2000) investigated students' conceptions of sampling. Participants included 62 students from third, sixth, and ninth grade in Australia. Conducting a qualitative analysis of students' written responses concerning sampling, Watson and Moritz (2000) identified three levels of thinking. At the first level, called unistructural, students possessed a primitive notion of sampling, referring to only a single aspect of a sample, such as a little bit. The second level, called multi-structural, combined several aspects. For example, a typical response may be, "a little of something, not the whole thing but a little piece of it" (Watson & Moritz, 2000, p. 53). Finally, the third stage, called relational, was obtained when responses indicated an understanding that a sample was a representative part from a larger whole. Understanding sampling is foundational for hypothesis testing. To draw inferences about a population, samples must

be unbiased, random, and drawn from the population of interest (Smith, 2008). However, the idea of taking a sample to draw a conclusion or make a decision can be problematic. Students often do not trust methods that yield representative samples and will select biased methods instead (Watson, 2004). There is also the problem of either thinking that a sample gives you no valuable information about the population or thinking that the sample tells you everything without room for error (Harradine et al., 2011).

Not only is the method of the data collection important, but the size of the sample is also critical in terms of the variability of the sampling distribution and the ability to achieve statistical significance. If a large sample size is used, one can detect differences even if they are small. However, if a small difference is not important, a conclusion of statistical significance may be misleading, because the results may not be practically significant.

Using simulations, Saldanha and McAllister (2014) investigated both important ideas in this section of determining what type of inferences can be drawn from a sample and the concept that increasing sample size reduces variability. Ninth graders engaged in three 65-80-minute lessons using Tinkerplots software to simulate resampling. The first part of the instructional sequence focused on data analysis and the second phase was devoted to inference. During the second phase, instructors focused on the idea that obtaining a representative sample allowed one to make conclusions regarding the population from which the sample was obtained. Additionally, students investigated the effect of increasing the sample size on the sampling distribution. The authors concluded that class discussions were essential in helping students understand key concepts. However, some students were hindered in selecting appropriate sample sizes to draw

conclusions from their inability to keep track of the three levels of sampling (population, sample, and sampling distribution), which will be discussed in more detail under the sampling distribution section. The basics of understanding the importance of how to collect the sample and what sample size is needed can be confounded by the more abstract ideas presented by the simulations. Saldanha and McAllister (2014) found that students who could not understand these abstract concepts of distribution of a sample and sampling distribution also struggled with selecting an appropriate sample size and determining a level of confidence.

Although it is important to understand that data collected randomly and without bias allows one to infer something about the population or cause and effect, students must not develop an overreliance on sampling representativeness. Understanding inference involves finding a balance between representativeness and variability (Shaughnessy, 2007). I will discuss the concept of variability and sampling distribution next.

Variability and Sampling Distribution

Cobb and Moore (1997) claimed that the field of statistics was created to deal with the omnipresence of variability. The importance of this concept was also emphasized in the GAISE report, which included a focus on variability among each problem-solving component (Franklin et al., 2007). Under the formulating questions section, one is said to anticipate variability. While collecting data, variability is to be acknowledged. Under analyzing data, variability is accounted for; and, finally, when interpreting results, one must allow for variability. Statisticians notice, acknowledge, measure, model, explain, and may attempt to control variability (Wild, & Pfannkuch, 1999). The existence of variability in this variety of contexts in statistics can make this

topic very difficult for students (Garfield, delMas, & Chance, 2007). For hypothesis testing, the context for variability is sampling. The idea is to seek to explain the variability that can be accounted for by chance alone. To do this, a model is needed to represent expected outcomes.

This model is called the sampling distribution of the null hypothesis. "The sampling distribution of a sample statistic describes how the value of the statistic varies from sample to sample" (Peck et al., 2013, p. 32). In the dolphin activity, the associated sampling distribution is based on randomization to treatments versus sampling from a population of interest. Therefore, for experiments, the sampling distribution is often referred to as a randomization distribution (Starnes et al., 2010); however, the underlying concept is still the same. The sampling or randomization distribution is used to see what types of results are expected when the null hypothesis is true. For the dolphin example, the null hypothesis was that the proportion of improvers would be the same for the dolphin and control group. In theory, the randomization distribution would have a mean of zero and the variability of the sampling distribution would be directly related to the group size.

In relating variability and sampling distribution to hypothesis testing, an essential understanding is recognizing the difference between the multiple levels of population, sample, and sampling distribution and the associated variability of each (Saldanha & Thompson, 2002). This was illustrated by the experiment that Saldanha & Thompson (2002) conducted with 11^{th} and 12^{th} grade students who were enrolled in a semester-long non-AP Statistics class. They investigated the ideas and interconnections between repeated sampling, variability among sample statistics, and distribution. Using

simulations, students were asked to focus on a three-level process which developed a simulated sampling distribution, like those created by the tasks in this study. In Level One, students would randomly select a sample of a particular size and record a statistic of interest. In Level Two, students would repeat this process many times until an accumulation of statistics were obtained. Finally, in Level Three, students would determine the proportion of statistics which were beyond a given value. This third level corresponds to obtaining a *p*-value and comparing it to an alpha-level in a traditional hypothesis test. Students who were able to correctly interpret the results of the simulation were those who could keep track of the differences among the multiple levels. They did not confuse the number of people in the sample with the number of samples taken, while still being able to coordinate the information as a whole. This allowed them to interpret the sampling distribution as a collection of sample statistics. More information regarding the specifics on the simulations used in the study conducted by Saldanha & Thompson (2002) are given under the simulations for hypothesis testing section in this chapter.

Saldanha and McAllister (2014) reported similar findings in their experiment, which used simulations to create sampling distributions with 9th grade students. The major goals of the activities were to develop the idea that random sampling can be used to draw conclusions and that larger samples reduce variability. To accomplish these goals, the activities focused on the variability of a simulated sampling distribution created from repeated sampling of a given size and by comparing that variability to sampling distributions created using different sample sizes. Students who struggled with correct conclusions often referred to the data points in the sampling distribution as individual values versus a statistic. Post-interviews reveled that these same students had difficulty

tracking the multi-tiered re-sampling process, like the struggles which students had in Saldanha and Thompson's (2002) study.

In both experiments described above, simulations were used to help develop students' reasoning concerning sampling distributions. Although positive changes did occur in many, some prevalent misconceptions still existed. Chance, delMas, and Garfield (2004) conducted a series of experiments using simulations to enhance students' understanding of sampling distribution in introductory college-level courses. They found that adding a conceptual change approach could enhance the use of simulations. With this instructional design, students were asked to answer questions regarding sampling distributions, then use computer software to test their claims. After comparing the simulation results with their answer, students were asked to reassess their conclusions. This approach resulted in statistically significant improvements in students' understanding. However, difficulties still existed with some students. Interviews revealed that students who still struggled often lacked the prerequisite vocabulary and understanding of distribution, variability, and models. Additionally, they found that some students were able to give correct answers due to memorizing rules and procedures.

In conclusion, the logic of hypothesis testing, probability, data collection, variability, and sampling distributions are all essential concepts of hypothesis testing. However, as indicated by research conducted with students, these are all difficult concepts to master. Many students often memorize the steps of a hypothesis test and are unable to articulate the reasoning behind these steps (Harradine et al., 2011). In the next section, I will relate what the literature reveals about teachers' understanding of hypothesis testing.

Teachers' Understanding of Hypothesis Testing

Few studies have explicitly studied teachers' understanding of hypothesis testing. In this section, I will relate the findings from studies which assessed teachers' understanding of hypothesis testing and related concepts. The first study investigated teachers' understanding of variability through comparing groups. Although traditional hypothesis testing was not investigated, this study did provide insight into teachers' understanding of informal hypothesis testing by focusing on interpreting variability and simulated sampling distributions. I will also discuss a study which investigated high school mathematics teachers' understanding of hypothesis testing as they were engaged in a professional development seminar.

Makar and Confrey (2004) conducted a qualitative analysis of four mathematics teachers engaged in a six-month professional development project. Their goal was to improve statistical content knowledge using activities and simulations. One participant was a pre-service teacher, and the other three were experienced teachers who taught 13-16-year old students. This study is particularly relevant to mine because one of the activities used involved simulating a null hypothesis sampling distribution to draw informal inferences. Additionally, in the final interview, participants were asked to complete a task in which simulated sampling distributions could be used as evidence to support their conclusions regarding group differences.

The researchers asked the participants to determine if two groups were different in terms of their achievement test scores and to provide evidence for their decision. Participants were very comfortable discussing the basic descriptive statistics, and they all addressed the variability within each data set. However, the participants struggled to

clearly express and provide evidence which addressed the variability between the two groups. Three of the participants mentioned that a sampling distribution may be useful, but none of them were able to articulate how this could be accomplished. They could also not explain the difference between the variability of the data set and the variability in the related sampling distribution. Additionally, none of the teachers accounted for the role of sample size and its effect on variability. The authors concluded that the concept of sampling distributions must be introduced more slowly and developed over time. This aligns with research reported in the previous section on the difficulty involved in understanding sampling distributions and variability for students.

Makar and Confrey (2004) did not report what classes their participants taught. In fact, one was a pre-service teacher and the others taught students from ages 13-16 years old. Typically, hypothesis testing is not encountered until 12^{th} grade. Therefore, it is likely that none of the participants had taught formal hypothesis testing. The next study I will discuss was reported by both Liu and Thompson (2009) and Thompson et al. (2007) and focused on high school mathematics teachers who had taught either stand-alone statistics courses or stand-alone statistics units, making their participants more like mine.

In Makar and Confrey's (2004) study, the participants engaged in a two-week professional development seminar, which focused on issues of inference and probability. The researchers reported that, like students, mathematics teachers do not understand the indirect reasoning behind hypothesis testing (Liu and Thompson, 2009). Thompson et al. (2007) reported that mathematics teachers used the logic that the null hypothesis was what they believed was true instead of focusing on trying to have evidence for the alternative. This is like the confirmation bias reported at the beginning of this chapter, in

which people look for evidence that will confirm their beliefs versus looking for evidence to disprove something.

One activity, which I also used in the pre-and post-interview with my participants, illustrated this lack of understanding of the logic behind hypothesis testing. For this scenario, participants were asked if they believed the claim that people prefer Pepsi over Coca Cola based on evidence from a sample in which 60% preferred Pepsi. They were also provided a simulated sampling distribution of sample proportions under the null hypothesis that the population was split 50-50 in their soda preference. In the simulation, a proportion of 60% or higher would occur only 2.96% of the time. Using the logic of hypothesis testing, one would reject the null hypothesis and conclude that people from the sampled population do prefer Pepsi more, which most participants did not conclude.

Based on the participants' discussion, Thompson et al. (2007) created a framework for teachers' logic of hypothesis testing (see Figure 2). Under decision one, the outcome is not unusual, and therefore the null hypothesis is not rejected. However, for decisions two and three, the outcome is unusual, and the null hypothesis should be rejected. Instead, for outcome two, the decision is to not reject the null hypothesis because the sample must have been biased. For decision three, even though the outcome was unusual, someone employing this logic believes that if the outcome had any possibility of occurring then the null should not be rejected. Finally, the fourth decision, which employs the logic of testing of providing evidence for the alternative, rejects the null hypothesis and concludes the alternative (denoted h_1 instead of h_a).

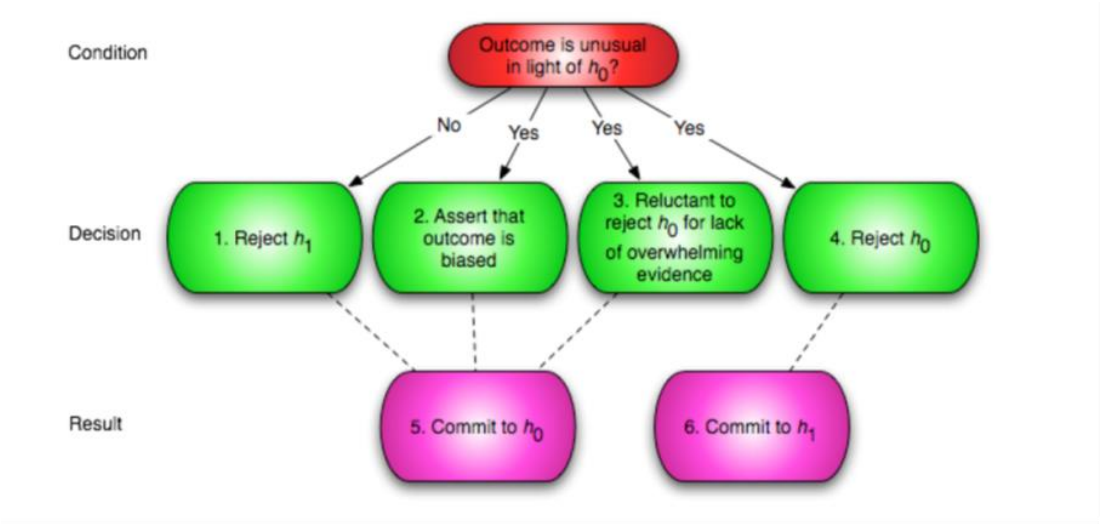


Figure 2. Theoretical Framework for teachers' logic of hypothesis testing (Thompson et al., 2007).

Thompson et al. (2007) also found that mathematics teachers did not appreciate scenarios in which hypothesis testing would be ideal to use. For example, when given a typical textbook scenario, only one of the participants suggested using a hypothesis test to test the validity of the claim. The others gave recommendations such as increasing the sample size, collecting more samples, or trying to sample the entire population.

Additionally, Thompson et al.'s (2007) reported that teachers had trouble understanding the difference between the distribution of averages of a sampling distribution and the distribution of individuals in a sample. This is like Saldanha and Thompson's (2002) study, mentioned under the variability and sampling distribution section and further discussed in the simulation section in this chapter, reported that students had difficulty understanding the difference between the three tiers of population, sample, and sampling distribution. Being able to understand this difference is thought to

be critical in interpreting results from a sampling distribution (Saldanha & McAllister, 2014; Saldanha & Thompson, 2002).

Understanding the role of probability is another important component of hypothesis testing which causes problems. Thompson et al. (2007) found that mathematics teachers thought of the idea of unusualness subjectively instead of stochastically. One participant claimed that something is unusual if it is unexpected and that those expectations are based on personal experience. All other participants, except for one, did not refer to the sampling distribution to quantify what was unusual. They did not understand that the *p*-value should be used to make a decision.

In conclusion, teachers often possess a similar understanding of hypothesis testing as students (Harradine et al., 2011). The studies reported in this section showed that teachers often do not understand the logic of hypothesis testing and struggle with some of the foundational concepts such as sampling distribution, variability, and probability. In the next section, I will discuss the research related to simulations for hypothesis testing.

Simulations for Hypothesis Testing

In general, students and teachers can often learn the associated computations and procedures of a hypothesis test but struggle with the reasoning and basic concepts of this procedure (Harradine et al., 2011). Even with intense remediation efforts, it is often difficult to convey an understanding of these abstract ideas (Thompson et al., 2007). Using simulations for hypothesis testing is a growing trend in statistics education due to the belief that simulations help make the abstract ideas behind hypothesis testing become more concrete and understandable (Erickson, 2006). This section will explain the

background of using simulations for inferences and share the knowledge gained from related studies.

What is Simulations for Hypothesis Testing?

Simulations for inference has its background in informal statistical inference (ISI). Recently, statistics education has included this idea to allow younger students to begin drawing inferences from data as early as elementary school, to prepare students to use more formal techniques later (Makar and Rubin, 2009). Makar and Rubin (2009) identified three key principles essential to ISI, which are: generalize beyond the data, use the data as evidence, and articulate uncertainty using probabilistic language. Generalizing beyond the data refers to the recognition that the data can tell you something about a larger population of interest. The actual data collected is then used to make your argument about what you believe to be true about that population. Finally, ISI also incorporates probabilistic language to articulate the uncertainty of conclusions. These three principles are also present in the simulations for inference tasks used in this study. However, a more sophisticated approach, moving towards formal techniques, is used. For example, the language of hypothesis testing was incorporated. Additionally, participants were asked to establish a null and alternative hypothesis and report a simulated *p*-value when drawing a conclusion. However, like ISI, formulas and theoretical sampling distributions were not used.

As mentioned in Chapter One, some introductory courses have moved to using simulation approaches exclusively to conduct hypothesis tests (Tintle et al., 2011; Tintle et. al., 2014). Recent research has indicated that students in classes using only simulations outperform students in classes using traditional approaches on assessments

aimed at measuring a conceptual understanding of statistical topics (Tintle et al., 2014). However, the focus of this study is on how using simulations for hypothesis testing can enhance teachers' understanding of traditional hypothesis testing.

This approach of using simulations to introduce hypothesis testing was used by Lane-Getaz and Zieffler (2006), who created a three-tier simulation process model (SPM) (see Figure 1). This model focused on addressing Saldanha and Thompson's (2002) claim that students must differentiate between the population, sample, and sampling distribution to understand inference. Additionally, the creation of the SPM was informed by a modeling approach, which makes explicit the connection between the concepts and models in activities (see Lesh, Cramer, Doerr, Post, and Zawojewski, 2003). In tier one, the population parameters are established, which is like forming the null and alternative hypothesis. Tier two involves generating sample statistics through simulation. This step replaces the traditional approach of using formulas to determine a test statistic and *p*-value. The final tier compiles the summary statistics to create an empirical sampling distribution to assess the unusualness of the observed data, which corresponds to using the *p*-value to reject or fail to reject the null hypothesis. Lane-Getaz and Zieffler (2006) recommended using this model to help students make sense of the simulation and make connections to traditional approaches. For this study, teachers will be asked to create their own simulation model, and I will look for comparisons.

Research about Simulations

I have selected studies in this section which specifically focus on using simulations to help students develop a deeper understanding of inference. A commonality among the studies is that each uses some type of model to illustrate the approach. This is

also relevant to this study because I will have participants create their own model of the simulation process and look at how they connect this approach to traditional hypothesis testing. Most like this study is Lane-Getaz and Zieffler's (2006) research, which focused on using simulations to develop a conceptual understanding of hypothesis testing in students and focused on connecting these two approaches. Additionally, their simulation approach is like mine in that they incorporate the recommendation of Rossman and Chance (2006) to use hands-on activities before the computer simulation to assist the learner in understanding how the computer is representing the distribution. Their work was also influenced by Saldanha and Thompson's (2002) recommendation to differentiate among the three levels of population, samples, and distribution of sample statistics. Therefore, I will discuss Saldanha and Thompson's (2002) study first, followed by Lane-Getaz and Zieffler (2006). Next, I will discuss Garfield, delMas, and Zieffler (2012) study which used a similar three-tier modeling approach. Although all three of these studies focused on simulation models which attended to three levels, Lee et al. (2016) recommended expanding on these levels. I will conclude this section with information about their study. As I investigate the teachers' models in this study, it will be important to attend to these possible expansions.

Saldanha and Thompson's (2002) study focused on how students conceptualized sample when using simulations, but they also connected their research to inference. Participants included non-AP high school statistics students. The researchers conducted a 9-session teaching experiment aimed at developing ideas of sample, sampling distributions, and margin of error. The use of computer simulations was an integral part of the teaching experiment. Researchers asked the students to focus on a three-step

process. First, students would randomly select a sample and calculate a statistic. Next, students would repeat this process many times. Finally, they asked the students to determine the proportion of statistics that were beyond a given threshold. This is like Cobb's (2007) three R's of randomize, repeat, and reject. Although not explicitly stated, the students were using simulations for hypothesis testing by creating a simulated sampling distribution to draw conclusions from data. Through examining the data collected from classroom discussions, written work, and interviews, the researchers noted two types of reasoning concerning sampling that emerged in their students. The first type of reasoning, which the researchers referred to as additive, involved seeing multiple samples as simply multiple subsets of the population. The students seemed to think that the subsets were being added to the sampling distribution, which led them to interpret the sampling distributions results as percent of people instead of percent of sample statistics. The researchers described students displaying a more sophisticated understanding as possessing a multiplicative conception of sampling. These students correctly interpreted the simulation results as a percent of sample proportions and seemed to understand the difference between the three levels of population, sample, and the distribution of the sample statistics. The following figure (see Figure 3) illustrates the multiplicative conception of sampling arising from the simulations. This was an important aspect that influenced the three-tier simulation approach adopted by Lane-Getaz and Zieffler (2006), which I will discuss next.

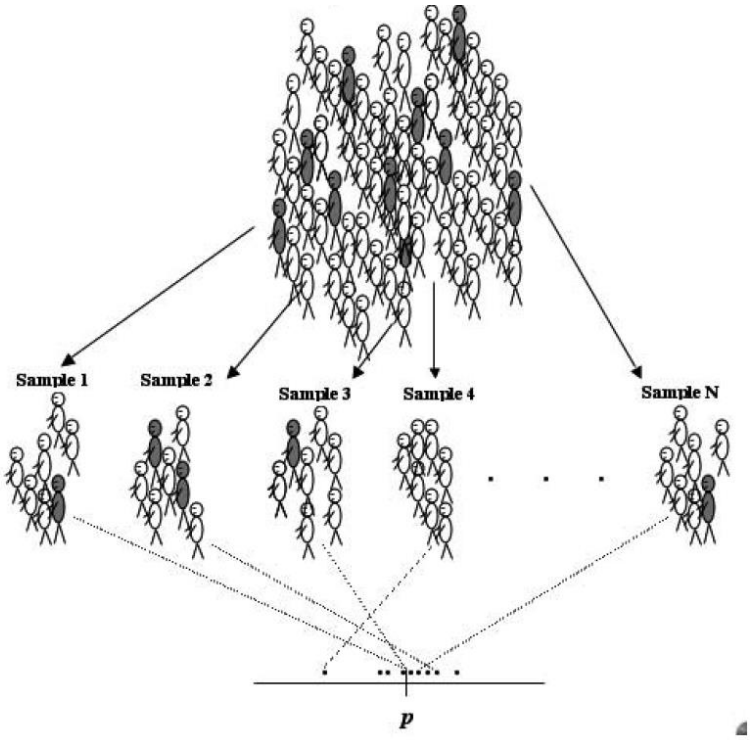


Figure 3. Three-Tier Multiplicative Conception of Sampling (Saldhana & Thompson, 2002).

Lane-Getaz and Zieffler (2006) investigated the use of what they referred to as the simulation process model (SPM) (See Figure 1). The SPM is a three-tier model, which they used to help students develop a conceptual understanding of hypothesis testing. The first tier referred to the hypothesized population, which corresponded to the null hypothesis in traditional hypothesis testing. The second tier corresponded to samples obtained and statistics calculated through simulation from the hypothesized population. The final tier was the compilation of those statistics, which resulted in an empirically derived sampling distribution. The SPM also included five steps, which aligned with the logic of inference. The first step referred to thinking about what if some model was true.

Step two was to generate samples, and step three was to select a statistic. In the fourth step, summary statistics are compiled. Finally, the last step assessed the rareness of the observed statistics. Students were exposed to the SPM before, during, and after being introduced to traditional hypothesis testing. During the second day of class, students investigated a claim and the resulting simulation was directly linked to this model by the instructor. The students seemed to understand the simulation but may not have understood all components of the model. Three more simulation activities were incorporated on days 5, 15, and 21. For each activity, students filled out SPMs. The instructor used days 21-25 to connect the SPM to traditional hypothesis testing. The diagram below provides an example connecting a two-sample *t* test for means to the simulation approach (see Figure 4).

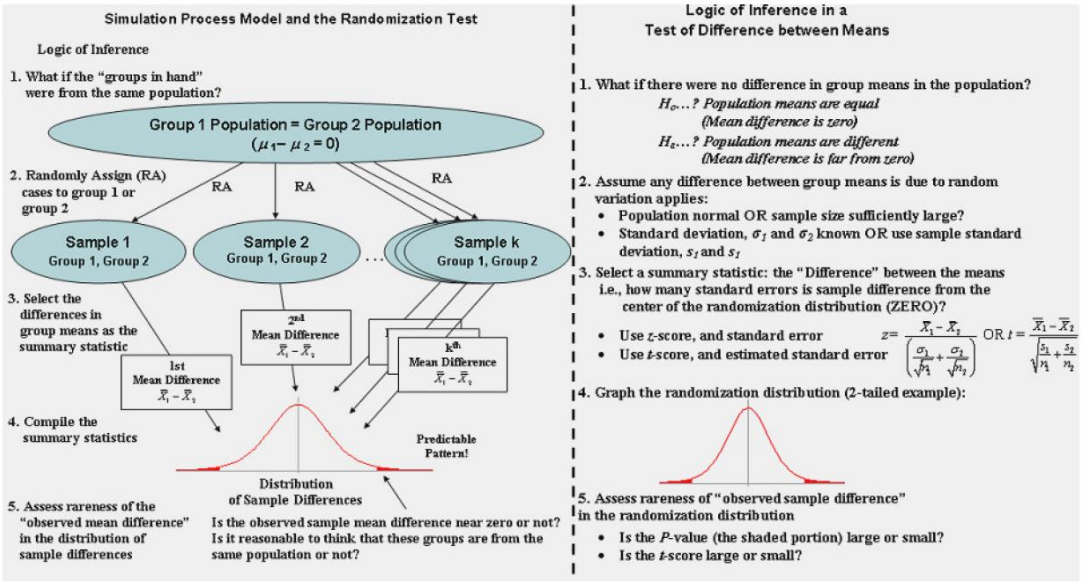


Figure 4. Connection of Simulation Approach to Two Sample *t* Test for Means (Lane-Getaz & Zieffler, 2006).

When connecting the approaches, Lane-Getaz and Zieffler (2006) used the five steps corresponding to what they referred to as the logic of inference instead of aligning the three-tiers from the SPM. Step one, which referred to establishing a hypothesized population, corresponded to step one of a traditional hypothesis test of establishing a null and alternative hypothesis. Step two, involved generating random samples through simulation. However, for traditional hypothesis testing, certain conditions must be checked, depending on the type of test. For step three, a summary statistic is selected for the simulation, and a test statistic is calculated under a hypothesis test. The fourth step involves compiling summary statistics for simulations, which is like the theoretical sampling distribution in hypothesis testing. The fifth step is similar for both and involves assessing the rareness of the observed results.

Lane-Getaz and Zieffler (2006) claimed that this approach allowed students to see the overall big picture behind hypothesis testing. However, they did not offer any specific evidence or examples. They did acknowledge that more research was needed to determine how this approach may help students think statistically. Next, I will describe a study which also used a similar modeling approach for simulations and compared students' conceptual understanding of statistics to students enrolled in classes using traditional inference procedures.

Garfield, delMas, and Zieffler, (2012) investigated the use of Change Agents for Teaching and Learning Statistics (CATALST) curriculum at the college level. The curriculum was inspired by Cobb and focused on using simulation methods for inference with the use of Tinkerplots. This method was based on a modeling approach of creating a

model, simulating data from the model, and drawing inferences from the simulated data. The figure (see Figure 5) below illustrates the three-tiered simulation approach used in this study. Students were solely exposed to simulations for inference and were not taught traditional methods.

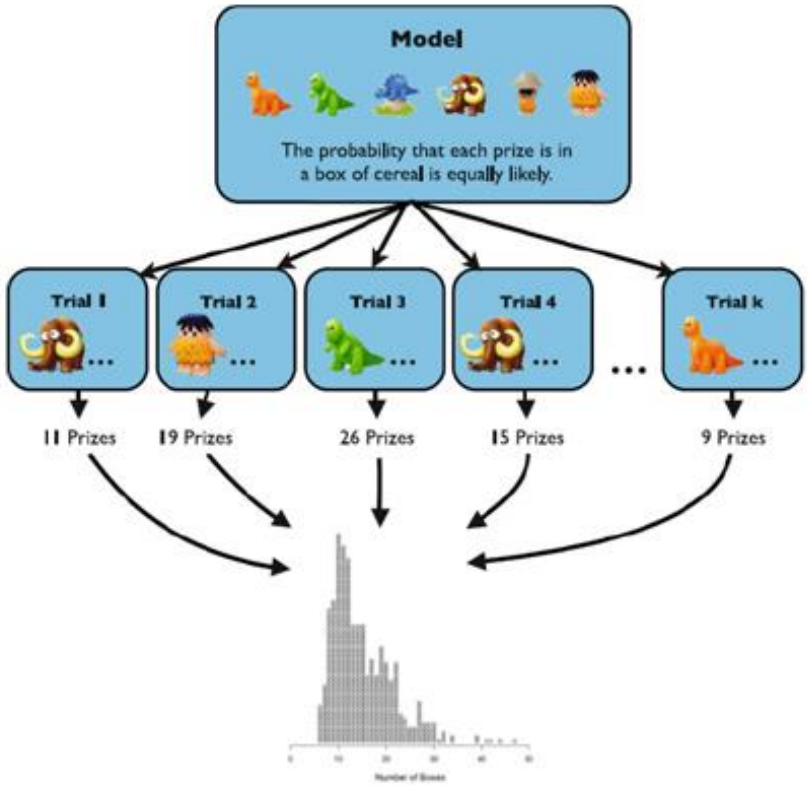


Figure 5. Example of Modeling and Simulation Process from CATALYST Curriculum (Garfield et al., 2012).

The study addressed four research questions. The first question concerned students' perception of using Tinkerplots. The second and third question focused on how

students learned to reason and think statistically using the curriculum. The final question focused on how the students felt about the course and statistics. The researchers used three instruments to answer their research questions. An affective survey was used to assess the students' attitudes and perceptions of Tinkerplots, the course, and the value of statistics. The Goals and Outcomes Associated with Learning Statistics (GOALS) assessment was used to evaluate the students' reasoning and basic understanding of statistical topics. The third instrument, Models of Statistical Thinking (MOST), was used to assess the students' statistical thinking. The researchers found that students had positive attitudes towards the use of Tinkerplots and did value statistics. Additionally, they found that students were beginning to think statistically and performed as well or better than students taking traditional statistics courses, even though not all the topics were taught explicitly. This study provided evidence that the use of simulations may help students better understand hypothesis testing in comparison to using traditional methods. Next, I will discuss a study using a similar simulation approach, but which also focused on how participants make connections to traditional approaches.

Lee et al. (2016) investigated the use of simulations for inference by students enrolled in a graduate-level course. Students in the class included one undergraduate pre-service teacher, 11 in-service teachers enrolled in a master's program, one full time masters mathematics education student, and eight doctoral students in mathematics or mathematics education. The classroom instructional design was based on a model development sequence. This sequence was comprised of seven activities geared towards helping the graduate students construct models and use simulations to conduct inference. These activities required the participants to describe results from a repeatable action and

identify an outcome of interest. For example, the Paul the Octopus scenario was used for one of the activities. For this scenario, Paul the Octopus correctly choose eight of eight winners for games by swimming to receive his food with the winning country's flag. The simulation investigates if Paul could really predict the winners or if he could have picked all eight winners just by chance. One of the other modeling eliciting tasks was based on the same dolphin treatment of depression scenario that was used in the third task of this study. One of the final activities asked the participants to create a visual model, which would help students understand the simulation approach used in the previous activities. The researchers selected five of these diagrams and analyzed them qualitatively. From this analysis they made several recommendations for helping students in the future understand and use simulation approaches, which will be discussed next.

Previous work (Garfield et al., 2012; Lane-Getaz & Zieffler, 2006; Saldanha, & Thompson, 2002) only recognized three levels of a simulation approach. Lee et al. (2016) recommended the first level, population or model, should be broken into two parts to make the pieces of this step more explicit. First, the model should be stated in terms of the real-world context, and next the actual simulation process should mimic the repeatable action. Additionally, they recommended being clear about the importance of creating a simulated sampling distribution and viewing it as a probability model. Finally, clearly referring to the observed statistic and using the empirical sampling distribution to ascertain how likely the observed result or something more extreme is to occur are other components that should be made more explicit for the learner. Lee et al. (2016) created the following chart (see Table 2) to show the important conceptualizations and corresponding affordances that should be made explicit in any simulation model.

Table 2

Key Conceptions and Capabilities for Learners using Simulations for Inference (Lee et al., 2016).

Conceptualization	Capabilities this conception affords
Conceive of events in the real-world problem as a result from a repeatable action	• Identify the underlying probability model of the event of interest (what is repeatable?) • Consider what results would be considered usual, or what would be considered usual or "to be expected". • Express a usual expectation as a null hypothesis. • Specify the observed statistic and the statistic of interest that should be observed when each action is repeated.
Conceive of and create a method for simulating the repeated sampling process	• Identify the repeatable action that needs to be enacted. • Choose tool (physical or computer) and map the action in the real world to a simple repeatable process using the tool.
Conceive of repeated sampling as a way to generate simulated statistics	• Recognize the need to enact the process for selecting a random sample of same size *n* and record the statistic of interest. • Repeat the random sampling process *k* times (large number) and collect the statistic from each sample for event of interest.

Conceptualization	Capabilities this conception affords
Conceive of how collected statistics from repeated samples vary with respect to likelihood	• Build a distribution of the recorded statistics. • Notice what seems to be usual (typical, or more likely to occur), and what is unusual (unlikely to occur). • Locate the original observed statistic in the distribution and consider whether it was in a range of "likely to happen" or "unlikely to happen".
Conceive of the inferential decision as involving deciding if the observed statistic and those more extreme are explainable by chance	• Use proportional reasoning to evaluate the likelihood that the observed event, and those more extreme, happened under the random process used to generate repeated actions and simulated statistics.

In conclusion, the previous studies all posited that using simulations can help students develop a deeper understanding of hypothesis testing. All these studies were at the high school or college-level; although, Lee et al. (2016) did have some in-service teachers enrolled in the graduate class involved in their study. This study will seek to add to this knowledge by focusing on high school statistics teachers. Additionally, the studies in this section all used some type of model for the simulation approach. The first three studies (Garfield et al., 2012; Lane-Getaz & Zieffler, 2006; Saldanha, & Thompson, 2002) used similar three-tier approaches, but Lee et al. (2016) recommended expanding the three tiers. Lane-Getaz and Zieffler's (2006) models for simulations and connecting

simulation and traditional approaches were used to help develop the theoretical framework for this study, which will be described next.

Theoretical Framework

A major component of this study was investigating teachers' understanding of hypothesis testing. However, multiple definitions for what it means to understand a topic exist, and frameworks can be a useful guide in determining what to investigate (Simon, 2006). The theoretical framework that I used for this study was based on Groth's (2013) Statistical Knowledge for Teaching (SKT), with modifications made to align the framework specifically to hypothesis testing and which are based on the literature reviewed in this chapter. In this section, I will provide a brief overview of Groth's SKT framework and then describe the specific theoretical framework used in this study.

Statistical Knowledge for Teaching (SKT)

In mathematics, Ball et al.'s (2004) Mathematical Knowledge for Teaching (MKT) is a well-known framework used to guide the analysis of mathematics knowledge needed for teaching. Groth (2007) was motivated to create a separate framework for statistics by the belief that statistics and mathematics are related yet distinct disciplines. Many statistical activities do require some mathematical reasoning; however, properties such as evaluating data in context and obtaining non-deterministic solutions are unique to statistics. Based on this premise, Groth (2007) stated that developing a separate framework for SKT was essential to foster appropriate teacher preparation and ongoing professional development. The original SKT framework used Hill et al.'s (2004) mathematical common and specialized knowledge, which will be explained later, as a basis, with a focus on the statistical problem-solving process described in the GAISE

report (Franklin et al., 2007). The four components of the problem-solving process included a) formulating questions, b) collecting data, c) analyzing data, and d) interpreting results. In the framework, the inclusion of both mathematical and nonmathematical knowledge under common and specialized knowledge was made explicit through each of the four statistical-problem solving components.

Groth continued to develop this model and elaborated the SKT framework in 2013. This model added contributions from Silverman and Thompson (2008) and Simon (2006) with the addition of Key Developmental Understandings (KDUs). KDUs are used to identify conceptual learning goals in mathematics (Simon, 2006). However, Simon (2006) stressed that this is not knowledge that can simply be explained or demonstrated to the student. Instead, KDUs represent a conceptual advancement which must be developed over time and which allow the student to perceive the underlying mathematical relationships.

Groth (2013) used this idea of KDUs as an anchor for portraying the specific components of subject matter knowledge and pedagogical content knowledge in statistics to develop his SKT framework (see Figure 6).

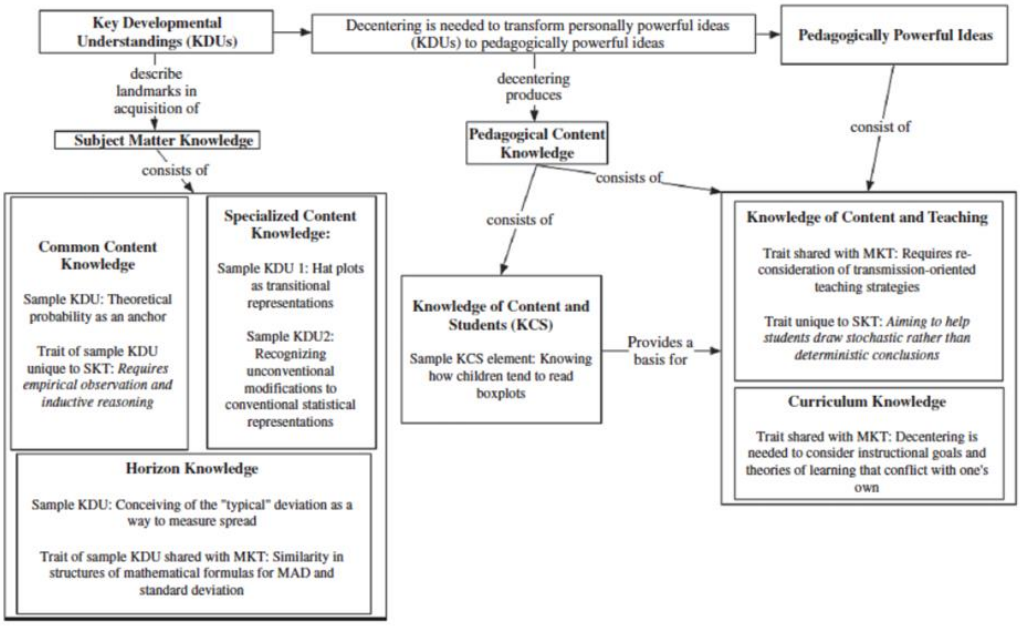


Figure 6. Hypothetical SKT elements and developmental structure (Groth, 2013)

The left-hand side depicts subject matter knowledge being comprised of Common Content Knowledge (CCK), Specialized Content Knowledge (SCK), and Horizon Knowledge. These components are based on Ball et al.'s (2008) Mathematical Knowledge for Teaching (MKT) framework. CCK was described as the basic knowledge of the subject matter that is not specific to teaching. It involved being able to correctly work problems and use appropriate terms and notation. The content knowledge that is unique to teaching was referred to as SCK. This knowledge is typically only needed for teachers. It goes beyond a basic understanding of how to correctly work a problem. A teacher must also be able to analyze if different solution paths will work and understand why they work. Horizon knowledge was described as the understanding of how the mathematics is connected across the curriculum. The right-hand side of Groth's (2013)

SKT framework concerns pedagogical content knowledge. This type of knowledge is comprised of knowing how students are likely to interpret concepts, being aware of what problems they may have, developing proper sequencing, choosing effective examples, and being aware of resources. For each category of knowledge, KDUs would be listed for the specific topic. I will describe how I modified this framework next.

Theoretical Framework: Hypothesis Testing CCK & SCK

Using the SKT framework as my guide, I have developed a specific theoretical framework for hypothesis testing (see Figure 7). This study will only investigate CCK and SCK for hypothesis testing. Therefore, I have included only these two relevant portions from Groth's (2013) framework.

.

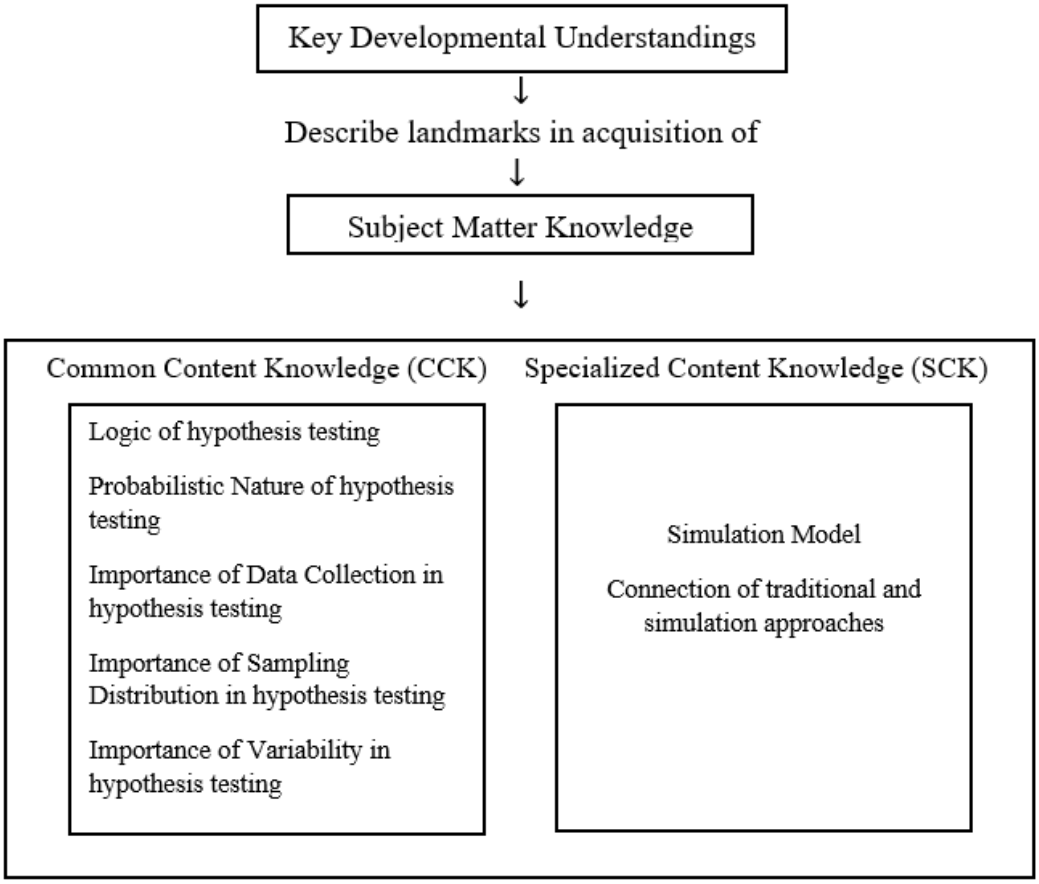


Figure 7. Theoretical Framework: Hypothesis Testing CCK & SCK adapted from Groth (2013)

My first research question relates to teachers' understanding of hypothesis testing that is not unique to teaching. Therefore, this type of knowledge corresponds to CCK. For my second research question, I looked at how teachers understand simulations from engaging in simulation tasks and how they make connections between simulation and traditional approaches. Understanding of simulations would be considered CCK if I had

investigated the subject matter knowledge of simulations. However, I investigated the subject matter knowledge of hypothesis testing and simulations from a pedagogical perspective of hypothesis testing. A simulation approach from a pedagogical viewpoint is a different solution path for a hypothesis test problem and is not required knowledge for understanding traditional hypothesis testing. Therefore, regarding traditional hypothesis testing, the knowledge of simulations would be mathematical knowledge typically only needed by the teacher, which would categorize this knowledge as SCK. Additionally, teachers need to be able to help their students see the connections between the simulation and traditional approach to deepen their students' understanding of traditional hypothesis testing. Therefore, this type of knowledge would also be categorized as SCK.

To determine the appropriate KDUs for this study, I focused on Simon's (2006) two characteristics of them. First, KDUs must involve a conceptual advancement in which the students' thinking and/or perception of mathematical relationships have been changed. Second, this type of understanding cannot be acquired through explanation or demonstration. Instead, multiple exposures to activities and reflections must be used. Next, I referenced the big ideas and essential understandings concerning hypothesis testing described in *Developing Essential Understandings of Statistics* (2003), mentioned at the beginning of Chapter Two, to guide my selection of KDUs for hypothesis testing. Thus, I have identified the KDUs for CCK of hypothesis testing as the logic of hypothesis testing, probabilistic nature of hypothesis testing, importance of data collection in hypothesis testing, importance of sampling distribution, and importance of variability. Additionally, teachers must understand the simulation approach and understand how to connect traditional and simulation approaches to facilitate a deeper understanding in their

students. However, this is not knowledge necessary to work the problem, but the knowledge needed to help students understand traditional hypothesis testing through simulations, which is an aspect of SCK. Therefore, for SCK, the relevant KDUs for this study are simulation model and connection of traditional and simulation approaches. These KDUs informed how I assessed my participants' subject matter of traditional hypothesis testing. Each KDU and how they were used in the research methodology are explained in detail under the data analysis and analytical framework section of Chapter Three.

Chapter Summary

In conclusion, in this chapter I provided an overview of the literature concerning hypothesis testing, which focused on five main concepts: logic of hypothesis testing, probability, data collection, variability, and sampling distribution. Additionally, I shared research regarding both students' and teachers' understanding of these topics. Next, I described simulations for hypothesis testing and the related literature. Finally, I concluded the chapter with a description of the theoretical framework, which guided this study. The KDUs from the theoretical framework directed the construction of the analytical framework used as part of the methodology of this study, which will be described next.

CHAPTER THREE: METHODOLOGY

Introduction

The purpose of this study was to investigate how high school statistics teachers' understanding of hypothesis testing is influenced by engaging in simulation tasks for hypothesis testing and to see how high school statistics teachers, who have not previously used simulation approaches, understand simulations and how they connect traditional and simulation approaches. Using an explanatory multiple-case study design, I sought to answer the following questions:

1. How does engaging in simulation tasks for hypothesis testing influence high school statistics teachers' understanding of traditional hypothesis testing?
2. How do simulation tasks influence high school statistics teachers' understanding of simulations and how do they make connections between traditional and simulation approaches for hypothesis testing?

In the first section of this chapter, I will provide an overview of the research methodology and describe myself as a research instrument. Next, I will describe participants and procedures, along with a procedure and timeline chart. I will then provide an explanation of the quantitative instrument used. Next, I will describe qualitative data sources, including an overview of the simulation tasks used in this study. I will then provide a detailed account of the data analysis, including the analytical framework. Finally, I will describe how quality and credibility were enhanced, along with the limitations and delimitations of the study.

Research Overview

I used an explanatory multiple-case study to answer my research questions of interest. An explanatory case study approach should be used when attempting to answer research questions involving how things have occurred (Yin, 2014). I have selected an explanatory approach, because I sought to explain how high school statistics teachers' understanding of hypothesis testing develops from the use of simulation tasks and to explain how simulation tasks influence high school statistics teachers' understanding of simulations and how they make connections between traditional and simulation approaches. I used multiple cases to establish validity of the results. I selected my participants from high school mathematics teachers who currently teach or who were preparing to teach AP Statistics or a similar non-AP Statistics class and had not previously used simulations for hypothesis testing. I selected three teachers with at least a general understanding of hypothesis testing so that they would be able to articulate their thinking processes concerning this topic. More details regarding the participants will be provided in the participants and procedures section in this chapter.

The purpose of using multiple cases is to serve as either literal or theoretical replications (Yin, 2014). A literal replication is used when the researcher is anticipating similar results, and theoretical replication is employed when contrasting results are expected (Yin, 2014). The three participants selected for this study were meant to serve as literal replications, which is like repeating an experiment in quantitative methodology (Yin, 2014). I had anticipated that these cases would be similar enough to serve as literal replications, because none of the participants had been exposed to simulations previously and each participant was engaged in the same tasks. This qualitative replication

strengthened my findings by providing multiple cases to analyze and cross reference (Yin, 2014). To triangulate data sources, I collected open-ended response questions, interview data, task responses, and post-task reflections. I will describe each of these in more detail under the procedures and qualitative data source sections. Additionally, I added one quantitative data source by giving participants inference portions of a pre- and post- CAOS test. Because inference refers to drawing a conclusion based on evidence and reasoning, which includes both confidence intervals and hypothesis testing, (delMas, 2004), I selected questions from the CAOS test which included not only hypothesis test questions, but also questions related to knowledge necessary for a deep understanding of a hypothesis test, such as sampling distribution and variability. I provide details regarding this test under the instrument section. With such a small sample size, I was not interested in statistically significant findings. However, I used this data as an additional source of information regarding the participants' understanding of inference.

Researcher

Qualitative designs use the researcher as an instrument in the data collection process (Patton, 2015). I engaged the participants in the tasks and conducted the interviews. Therefore, I will describe myself, my interest in this topic, and previous experiences with simulations to enhance the credibility of the study.

I have been a mathematics teacher for 17 years, and I have taught statistics for 15 of those years. Like other mathematics teachers, I felt completely unprepared to teach this subject. Additionally, I felt a sense of isolation and distinct lack of resources. However, I quickly fell in love with statistics and embraced this subject, which seemed so different from the other mathematics classes that I taught. Even the textbook, which was full of

words instead of numbers, seemed foreign to myself and students alike. However, I found it fulfilling to have my students say that they were finally in a math class that they felt was useful and pertinent to their lives.

I found it easy to make statistics relevant to my students, but I always found the inference topics to be the most difficult to teach. Unfortunately, for many of my students, I would just say, "Well, if you don't get it, just memorize the steps, and you'll get the answer right." This was not something that I was proud of, but I did not see another way to explain some of the more complicated details, such as what is a sampling distribution and what a *p*-value means. The result was that some students eventually developed a conceptual understanding of a hypothesis test, but many of them had just memorized the procedure of "if the *p* is low, reject the ho (null hypothesis)." This may have led to some laughs but not to the desired outcome of a rich understanding of a fundamental topic in statistics.

Upon entering my mathematics education PhD program and being exposed to reform-based practices, I took a hard look at many of my instructional practices. At the same time, my faculty mentor introduced me to the concept of using simulations for inference. Of course, my statistics students still had to know all the steps of a traditional test to correctly answer questions on the AP exam that they would take. However, I saw this as an opportunity to develop their conceptual understanding before their procedural fluency. I taught a couple of lessons that year and noticed that my students were no longer confused about if they should reject the null hypothesis if the *p*-value was large or small, which was evidence to me that they were developing a deeper understanding of

this topic. Although anecdotal, I truly believe that the use of simulations was what led to this outcome.

Additionally, I found that my own content knowledge was strengthened using these approaches. I realized that I had been incorrectly describing the *p*-value as a way to determine which hypothesis was more likely correct. I also found that my ability to explain the concept of a hypothesis test to both students and colleagues had developed. After a presentation on this topic, I had a fellow student in my doctoral program tell me that they had never understood the logic behind a hypothesis test until I had described it with the aid of simulations.

These experiences have convinced me that using simulations for inferences is an ideal way to develop a deeper understanding of hypothesis testing for both students and teachers. Although the advancement of technology has made the use of traditional approaches unnecessary, they are still part of most introductory statistics classes (Cobb, 2007). Without a way to introduce this topic conceptually, students and teachers may continue to focus on procedures, rather than the underlying logic behind a hypothesis test. With these beliefs transparent, I will be sure to adhere to strict and established data collection and analysis procedures, as described under the data analysis section, to ensure the reliability of this study.

Participants and Procedures

I purposefully selected three current high school teachers in the southeastern United States who possessed at least a procedural knowledge of hypothesis testing but had not used simulation approaches previously. Possessing at least a procedural knowledge of hypothesis testing was important so that the participants had enough

knowledge to articulate their reasoning and thinking processes. I will use the pseudonyms Carrie, Kathleen, and Chase to refer to the participants.

At the time of selection, all three teachers were supposed to teach statistics during the school year that data collection was to take place. However, Chase's schedule changed, and he was no longer teaching this subject. Although Chase had not previously taught an AP or non-AP statistics class, he had received AP statistics training and tutored statistics. Additionally, observing the teachers in the classroom was not part of the data collection process. Therefore, Chase could be considered a future statistics teacher, and he was kept in the study.

Chase participated in all data collection phases and the tasks. However, in the post-open-ended response survey, Chase revealed that he had engaged in simulation tasks for hypothesis testing before this study in an AP training. The trainer called this approach using simulations instead of using simulations for hypothesis testing, as I had called it when explaining the study to Chase. Therefore, Chase did not realize he did not meet the selection criteria and was mistakenly kept in the study. Due to his knowledge having been influenced previously by his training with simulations, he is not representative of my population of interest. Therefore, his data will not be reported in this dissertation. However, his dialogue was an important component when the other participants engaged in the tasks, so I have left information regarding him in this section.

Participants were compensated for their time with a $100 gift card and gained knowledge concerning using simulation approaches to incorporate in their classrooms. They also received copies of the lesson plans for the simulation tasks which were used in data collection. The data collection process included a pre-task phase, task phase, and post-

task phase. The pre-task phase was used to obtain an initial gauge of the participants' understanding of hypothesis testing. Participants were asked to complete a background information survey (Appendix A), pre-open-ended response questions (Appendix B), and the inference portion of the CAOS test (Appendix C). The CAOS test had items regarding the foundational topics of hypothesis testing, such as sampling variability and *p*-value (delMas et al., 2007). By obtaining an initial measure of these teachers' foundational understanding of these topics, I was able to develop more meaningful and probing pre-task interview questions.

Therefore, after I evaluated the results of the open-ended response questions and the CAOS test, I scheduled a meeting with each participant to conduct a pre-interview to further examine the participants' understanding of hypothesis testing. A semi-structured interview protocol (Appendix D) was used to obtain the remaining pre-task data, which was used to determine the participants' understanding of hypothesis testing before engaging in simulation tasks. During the interview, I made sure to ask specific follow-up questions from the pre-CAOS. For example, Carrie had missed most of the questions regarding *p*-value, so I focused on having her explain this concept.

Next, I collected data during the task phase. I scheduled three group meetings, each one approximately one week apart, to engage the participants in three tasks using simulation approaches, designed based on a conceptually-focused, six-phase lesson structure for using simulations for hypothesis testing (see Strayer & Matuszewski, 2016). The lesson plan for each task is described in detail under the qualitative data sources section in this chapter. Task one involved a one sample *z*-test for proportions. The following two tasks involved a two sample-test for means and proportions, respectively.

After each task, participants completed a reflection component individually (see Appendix E). The purpose of the post-task reflection was to gain insight into how the participants connected traditional and simulation approaches. After completing all three tasks, participants completed post open-ended questions (Appendix F) regarding their understanding of hypothesis testing and completed the post-CAOS test. A short initial analysis of the data was conducted to develop probing questions for the second interview portion of the study. A semi-structured interview protocol (Appendix G) was used to allow for emergent themes and provide relevant information for the research questions of interest. An overview of the data collection plan is provided in Figure 8 and a timeline is given in Table 3. Specific details regarding the analysis components mentioned in Figure 8 will be described under the data analysis section.

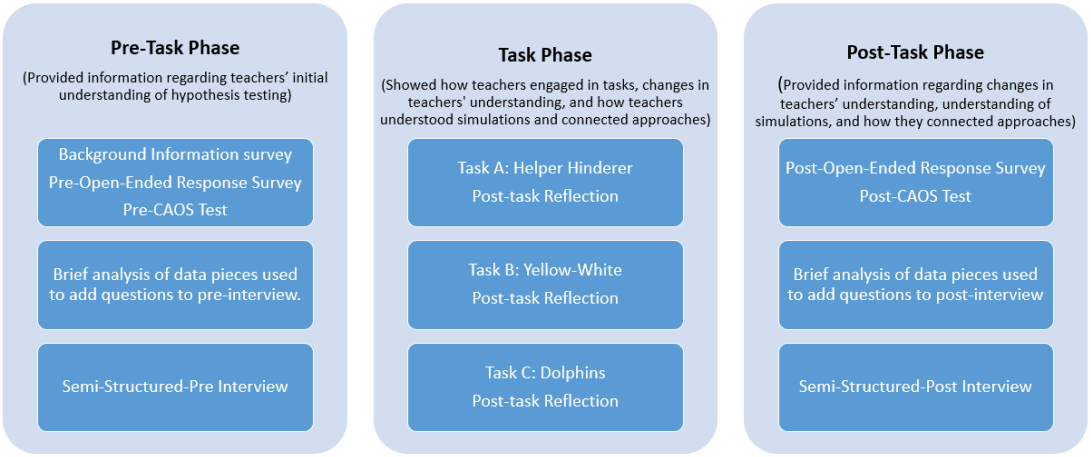


Figure 8. Overview of Data Collection Plan

Table 3

Timeline for Data Collection (2017)

Week	Data collection
October 2nd	Completed background information, pre-open-ended response questions, and inference portion of pre-CAOS test
October 9th	Brief analysis and added additional questions to semi-structured pre-interview protocol
October 16th	Pre-interviews
November 6th	Task A
November 13th	Task B
November 20th	Task C
November 27th	Completed post-open-ended response questions and post-CAOS test
December 4th	Brief analysis of data to add additional questions for post interview
December 18th	Post-interviews

Instrument

For the quantitative data source for the study, I had each participant take a pre- and post-test of the inference portions of the Comprehensive Assessment of Outcomes in

a First Statistics course (CAOS) test. The CAOS test was developed by the Assessment Resource Tools for Improving Statistical Thinking (ARTIST) project, which was funded by the National Science Foundation (NSF) (delMas et al., 2007). After three years of testing, the final version of CAOS consisted of 40 multiple choice items and was found to have an acceptable reliability rating of .82 (delMas et al., 2007). To obtain access to test items, one may request permission at https://apps3.cehd.umn.edu/artist/index.html. I only used the portions of the test applicable to inference, which included the subsections titled tests of significance and sampling variability. The purpose of the instrument was to offer an additional data source regarding participants' understanding of hypothesis testing and inference topics in general.

Qualitative Data Sources

Five sources of qualitative data were collected, including a background information survey, data from rounds of interviews, task data, post-task reflections, and open-ended responses. Participants completed a background information survey (Appendix A) and initial open-ended response questions (Appendix B) so that the researcher could develop a description of each case and gain insight into the participants' understanding of hypothesis testing. The first interview took place before the participants engaged in the simulation tasks. Using a semi-structured interview protocol (Appendix D), participants were asked questions to further elucidate their understanding of hypothesis testing. Next, each participant completed three tasks as a group, with myself facilitating and with approximately one week between each task. I video recorded each task and had participants record their responses on a handout. Participants were encouraged to think aloud while engaging in the task, and I transcribed the process. After

completing each task, the participants completed a reflection component (Appendix E). After completing all three tasks, the participants then completed open-ended questions (Appendix F) to assess their understanding. The final data source came from a post-interview, again using a semi-structured interview protocol (Appendix G). Next, I will provide a general overview of each lesson plan task. The specifics regarding how the participants interacted as they completed the tasks will be provided in Chapter Four.

Task A: Helper-Hinderer (Holcomb, Chance, Rossman, Tietjen, & Cobb, 2010)

This task introduces the concept of using simulations for inference with a one-sample *z*-test for proportions involving a null hypothesis of 50% to simplify the simulation model. The task scenario involved a study conducted by Yale researchers in 2007 in which groups of 6-month-olds and 10-month-olds watched a puppet show with neutral wooden figures, where one figure, the climber, was trying to get up a hill. In one scenario, one of the other figures, called the helper, assisted the climber up the hill. In the other scenario, a third figure, called the hinderer, pushed the climber down. Participants were asked if 16 pre-verbal children participated in the study, how many do they believe chose the helper toy. Next, they were asked to list the possible hypotheses for the study. The null hypothesis was that children have no preference, and participants then determined what the most expected result would be if the null hypothesis is true, which was 8, for a 50% chance if the children have no preference. Next, the participants determined the types of results which would not surprise them if the null hypothesis were true. After recording their own opinion of what would not be surprising, I revealed that 12 out of the 16 (75%) children preferred the helper toy in the original experiment. However, this led to the need for a simulation model to see what type of results would be

typical if the null hypothesis were true. Because the null hypothesis was no preference, meaning that 50% would prefer the helper toy, participants selected using a coin with heads representing a child being in favor of the helper toy and tails representing not being in favor to conduct the simulation. Additionally, because the sample size was 16, participants flipped the coin 16 times and determined the number of times the coin landed on heads to represent one trial. To see what types of results are typical and to obtain a more accurate estimation of the *p*-value, many trials would need to be conducted. Technology was used to simulate this many times and obtain an approximate *p*-value by seeing how many times a trial resulted in 12 or more selecting the helper toy. Participants were then asked to make a decision regarding whether or not children prefer the helper toy.

Task B: Yellow-White (Statcrunch, n.d.)

This task is like a hypothesis test for comparing two means. The context involves test scores obtained from the same test being given to students, but half of the students taking the test on yellow paper and half on white paper. The question of interest is whether students taking the test on the yellow paper, which some believe is a more peaceful color than the stark white, would perform better. Participants were asked if 20 students took the exam, 10 on white paper and 10 on yellow, how do they think the average score from students who took the exam on yellow paper would compare to average score of the students who took the exam on white paper and why. Next, participants were asked to determine the different hypothesis they could make regarding the averages of scores of students who take the exam on yellow paper and on white paper. The possible hypotheses are that the average scores are the same, the average score

for yellow paper is higher, and the average score for yellow paper is lower. The null hypothesis is that the averages are the same, and participants were asked to state what they believe would be the most likely outcome (difference in the average scores) when this study is conducted with 20 participants and the null hypothesis is true. If the average scores are the same, then a difference of zero would be the most likely outcome. Next, still assuming that the color of the exam did not affect students' scores (i.e. students would get the same score regardless of the color of the exam) participants were asked to relate what kind of results (difference in the average scores) would they not be surprised to see when this study is conducted with 20 participants. The next phase was the revelation of the actual study's results, which was that the difference in average scores was actually 6.3. Participants were then asked to design a simulation under the assumption that the null hypothesis is true. Slips of paper with the test scores were provided. Participants shuffled the test scores and dealt out 10 of the scores to a group representing the yellow exam paper and the remaining 10 representing the white paper. The participants calculated the average of both groups and subtracted them to complete one trial of the simulation. Next, technology was used to quickly simulate the scenario many times, and the participants were asked to make a conclusion based on the probability of obtaining a difference of 6.3 points or more.

Task C: Swimming with Dolphins (Rossman, 2008)

The context of the task was an experiment which analyzed the effectiveness of dolphin therapy to relieve mild to moderate depression conducted by Antonioli and Reveley (2005). In the experiment, 30 subjects were randomly assigned to one of two groups. In the control group, subjects swam and snorkeled each day. The treatment group

also swam and snorkeled each day, but they did so in the presence of bottlenose dolphins. At the end of the study, the subjects' level of depression was reevaluated to determine if they showed substantial improvement. I revealed part of the study's results that 13 of the 30 subjects showed improvement and had the participants commit to how many of the 13 they thought were in the dolphin group. Next, I had the participants list all the hypotheses for the situation. This resulted in three hypotheses: swimming with dolphins relieves depression, swimming with dolphins does not relieve depression, and swimming with dolphins has no effect on depression.

Next, I asked the participants what kind of results they would anticipate if there were really no differences in the groups, which was the null hypothesis. After a brief discussion, participants agreed that about half of the improvers should end up in each group if there is really no difference. However, they discussed that because there are 13 improvers, the most likely outcome would be that 6 or 7 improvers would end up in each group. Finally, I revealed the rest of the study's results that 10 improvers were in the dolphin group and 3 were in the control group. To determine if this was enough evidence to conclude that dolphin therapy is effective, participants needed to find out what types of results are likely if the null hypothesis is true. Before using technology to quickly simulate the types of results expected, I had participants determine how to simulate the experiment themselves. This required prompting participants to recall that possible simulation materials included coins, chips, a calculator, and cards. I questioned the participants on how they are mimicking the randomization process and what outcome of interest they would record. After some discussion, the participants chose to use cards. Thirteen red cards represented the improvers, and 17 black cards represented the non-

improvers. Next, the participants shuffled all 30 cards and dealt out 15 cards in two piles to mimic the randomization of participants to the treatment and control group. After that, they counted the number of improvers in each group and recorded the difference in number of improvers.

The purpose of the simulation was to get the participants to recognize that you are trying to determine what types of results are typical when you repeated the random assignment of subjects to the two treatment groups with only chance accounting for the variability in number of subjects showing substantial improvement in each group. To determine an approximate *p*-value, technology was used to simulate the experiment thousands of times. The participants then analyzed the approximate sampling distribution and made a conclusion regarding if dolphins help relieve depression.

Data Analysis and Analytical Frameworks

I began the data analysis process by organizing my data. The data included pre- and post-CAOS assessments, pre- and post-open-ended response surveys, pre- and post-interview data, task data, and post-task reflections. I had scored each participants' pre- and post-CAOS test before conducting their pre- and post-interview. Next, I organized this data in a table to show a comparison of which items they missed before and after the simulation tasks. Similarly, I created tables which compared their pre- and post-open-ended survey responses. For the pre- and post-interview and task data, I transcribed all the data myself and listened to the audio several times. This allowed me to begin the qualitative journey of immersing myself and becoming familiar with the data (Bloomberg & Volpe, 2012). After transcribing and organizing my data, I began my data analysis steps.

The following table (see Table 4) provides an overview of my data analysis steps, which was influenced by the systematic procedure for data analysis described by Bloomberg and Volpe (2012).

Table 4

Data Analysis Steps

Stage	Processes
One: Explore Data and Coding	• Rereading of all data pieces to obtain an overall sense of the whole and gain initial insights. Writing of analytical memos. • Deductive coding based on KDUs from theoretical framework used to identify data pieces corresponding to broad categories for CCK and SCK. • Inductive coding used to explore data and find factors which influenced changes in understandings.
Two: Analytical Framework and Narratives	• This framework divided each KDU from the theoretical framework into a list of understandings. • For different time periods, data sources were located and coded by understanding. These data pieces were assessed for understanding by comparing each data piece to the listed understanding in the analytical framework and writing analytical memos concerning the alignment of the data piece with the understanding. • Narratives written for pre-data, post-data, and post-task reflections based on categories of CCK and SCK.
Three: Review and Revise	• Iterative process used to reread all data pieces and check narratives for accuracy.
Four: Additional Narratives, Develop Themes, and Accuracy Check	• Narratives for tasks produced. • Inductive codes, analytical memos, and narratives used to determine themes. • Rereading and checking of all data pieces, coding, narratives, memos, and themes.
Five: Analyze Second Case	• Repeat steps two through four for second case. • Refining of overall themes for both cases.
Six: Cross Case Analysis	• Compared pre- and post-data, changes in CCK, understanding of simulations, connection of approaches, and themes.

In stage one, I began by reading all my data pieces and open coding portions of my body of data, which allowed me to conceptualize what was happening on the data's own terms through an inductive approach. As I was inductively coding to gain initial insights, I

simultaneously used deductive coding to identify pieces of data which revealed my participants' CCK and SCK of hypothesis testing. My list of deductive codes, which classified the data according to the general categories for the participants' CCK and SCK, was produced from the theoretical framework. Recall that the theoretical framework described CCK of hypothesis testing as being comprised of five KDUs. These KDUs, derived from the literature, were logic of hypothesis testing, probabilistic nature of hypothesis testing, importance of data collection in hypothesis testing, importance of sampling distribution, and importance of variability. For SCK, the KDUs were simulation model and connection of traditional and simulation approaches. The following table (see Table 5) shows the codes produced both deductively and inductively during stage one of the data analysis process.

Table 5

Initial List of Codes

Deductive Codes	Inductive Codes
Logic of hypothesis testing	Hypotheses
Probabilistic Nature	Alpha
Data collection	*P*-value
Sampling Distribution	Errors
Variability	Expected results
Simulation	Visualization
Connecting Approaches	Procedures
	Lesson Plan
	Context

Additionally, as I was going through this process, I also began crafting my initial analytical memos to obtain an overall sense of the story that my data was telling me. For example, in an analytical memo, I wrote that Carrie's response to the question which asked her to explain the logic of hypothesis testing on the pre-opened-response survey had focused on procedures and her post-open-ended response to the same question focused on concepts. Also, by comparing the pre- and post-data, I noticed that my participants' content knowledge had changed, and by reading the task data, it seemed that certain aspects of the tasks had caused these changes. For example, I noted that the visualization component of the tasks seemed to influence how the participants understood *p*-value.

Next, in stage two, I used the analytical framework to further assess the data. I created this framework prior to data collection, because I knew that I would need a list of specific understandings of hypothesis testing that I could use to assess my participants' CCK, which would be used to answer research question one. Also, I needed a model for simulations and for connecting approaches for assessing their SCK, which would be used to answer research question two. The analytical framework, based on the literature, provided this. This framework (see Table 6) was created by taking the five KDUs which corresponded to CCK and the two KDUs which corresponded to SCK from the study's theoretical framework and dividing them into a list of understandings. The first set of understandings listed, which corresponded to the five KDUs for CCK, were obtained from the understandings listed in Smith's (2008) Framework for Assessing Understanding of Statistical Hypothesis Testing. The understandings listed for the last two KDUs for SCK were derived from modified versions of Lane Getaz and Zieffler's (2006) SPM and connecting approaches model. I will explain in detail how I used this framework next.

Table 6

Analytical Framework for Assessing Understanding of Statistical Hypothesis Testing and Simulations Modified from Smith (2008) and Lane-Getaz and Zieffler (2006).

KDU	Understanding Assessed
Logic of hypothesis testing	Indirect reasoning will be employed and, therefore, two competing hypotheses are needed. Writing the hypothesis to indicate a one- or two-tailed test will address the practical needs of the researcher. Hypothesis testing provides a means of "answering" a research question about a population from a sample. Failing to reject the null hypothesis does not prove the null hypothesis. Statistical significance does not mean practical significance.
Probabilistic Nature	A "cut point" is necessary to determine whether to reject the null hypothesis or not. This decision considers the probability associated with the sample. A "cut point" determines the probability of a Type I error. The *p*-value is the probability of getting values as extreme or more extreme as the observed value, if the null was true. If the *p*-value is small, this means the result was unlikely if the null was true and the null should be rejected. You have evidence for the alternative. If not, you do not have evidence for the alternative.
Data Collection	Samples must be unbiased and random. Larger samples are more representative of the population. The way in which a sample is chosen will affect the nature of the inference that can be drawn, including the population to which the inference can be applied.

KDU	Understanding Assessed
Variability and Sampling Distribution	There is a difference between the variability of the population, the sample, and the sampling distribution. Samples are expected to vary. For a given sample size, n, and sample statistic, the sampling distribution of the statistic gives a probability distribution of values taken by the sample statistic for all possible samples of size n. It is not the distribution of a particular sample. The variability of the sampling distribution is influenced by the size of the sample. To determine if a sample is unusual, one should examine the sampling distribution of the given statistic, for samples of size n, under the assumption that the null hypothesis describes the population.
Simulations	The steps to conduct a simulation are: 1. Establish population parameters 2. Generate samples through simulation 3. Create sampling distribution and assess unusualness Connection of Approaches: Step1: Simulation: What if scenario? Determine a model. Hypothesis test: Statement of null and alternative hypothesis Step2: Simulation: Repeat the random sampling or assignment through simulation. Hypothesis test: Check conditions Step3: Simulation: Select the appropriate summary statistic Hypothesis Test: Calculate *z* or *t* test statistic Step4: Simulation: Compile summary statistic for distribution formed by simulation Hypothesis Test: Graph the theoretical sampling distribution based on a function Step5: Simulation: Assess rareness by finding observed data on simulated sampling distribution and calculating approximate *p*-value. Hypothesis Test: Assess rareness by large or small *p*-value OR large or small test statistic.

To help me use the analytical framework, I relied on a table that I had also constructed prior to data collection. This table (see Table 7) aligned each of my research questions to the corresponding KDUs, the list of understandings for each KDU from the analytical framework, and the related data sources. This allowed me to know which data pieces contained the information that I needed to assess to determine the participants' understanding for each KDU and which research question this data would be used to help answer.

Table 7

Alignment of Research Questions, KDUs, Analytical Framework, and Data Sources

Research Question	Theoretical Framework KDUs	Analytical Framework Understanding Assessed	Data Sources
Q1 How does engaging in simulation tasks for hypothesis testing influence high school statistics teachers' understanding of traditional hypothesis testing?	KDU: Logic of hypothesis testing	Indirect reasoning will be employed and, therefore, two competing hypotheses are needed.	Pre- and Post-Open-ended questions 3-5 Pre-interview questions 1-7 Post-interview questions 1-3 Tasks and Post-task reflections
		Writing the hypothesis to indicate a one- or two-tailed test will address the practical needs of the researcher.	CAOS Test of Sig 1 Pre- and Post-Open-ended questions 1,4-5 Pre-interview questions 1-5,7 Tasks and Post-task reflections
		Hypothesis testing provides a means of "answering" a research	Pre-and Post-Open-ended questions 1-5 Pre-interview questions 1-7

Research Question	Theoretical Framework KDUs	Analytical Framework Understanding Assessed	Data Sources
		question about a population from a sample.	Post-interview questions 1-3 Tasks and Post-task reflections
		Failing to reject the null hypothesis does not prove the null hypothesis.	Pre-and Post-Open-ended questions 3-5 Pre-interview questions 1-3, 6, 7 Post-interview questions 1-3 Tasks and Post-task reflections
		Statistical significance does not mean practical significance.	CAOS Test of Sig 7 Pre-interview questions 7 Post-interview questions 3 Tasks and Post-task reflections
Q1	KDU: Probabilistic Nature	A "cut point" is necessary to determine whether to reject the null hypothesis or not. This decision considers the probability associated with the sample.	Pre-and Post-Open-ended questions 2, 4, 5 Pre-interview questions 1, 7 Post-interview questions 1-3 Tasks and Post-task reflections
		A "cut point" determines the probability of a Type I error.	CAOS Test of Sig 4 Pre-and Post-Open-ended question 2 Pre-interview questions 7 Tasks and Post-task reflections
		The *p*-value is the probability of getting values as extreme or more extreme as the observed value, if the null was true.	CAOS Test of Sig 2 Pre-and Post-Open-ended question 2 Pre-interview questions 5, 7

Research Question	Theoretical Framework KDUs	Analytical Framework Understanding Assessed	Data Sources
			Post-interview questions 1-3 Tasks and Post-task reflections
		If the *p*-value is small, this means the result was unlikely if the null was true and the null should be rejected. You have evidence for the alternative. If not, you do not have evidence for the alternative.	CAOS Test of Sig 3, 10 Pre-and Post-2, 4, 5 Pre-interview questions 5, 7 Post-interview questions 1-3 Tasks and Post-task reflections
Q1	KDU: Data Collection	The way in which a sample is chosen will affect the nature of the inference that can be drawn, including the population to which the inference can be applied.	Pre-and Post-Open-ended questions 4, 5 Pre-interview questions 1-7 Post-interview questions 1-3 Tasks and Post-task reflections
		Samples must be unbiased and random.	CAOS Test of Sig 8 Pre-and Post-Open-ended questions 4, 5 Pre-interview questions 1-7 Post-interview questions 1-3 Tasks and Post-task reflections
		Larger samples are more representative of the population.	CAOS Test of Var 2,3 Pre-and Post-Open-ended questions 4, 5 Pre-interview questions 1-7 Post-interview questions 1-3

Research Question	Theoretical Framework KDUs	Analytical Framework Understanding Assessed	Data Sources
			Tasks and Post-task reflections
Q1	KDU: Sampling Distribution.	For a given sample size, n, and sample statistic, the sampling distribution of the statistic gives a probability distribution of values taken by the sample statistic for all possible samples of size n. It is not the distribution of a particular sample.	CAOS Test of Var 1 Pre-and Post-Open-ended questions 2 Pre-interview questions 2,3,5,7 Post-interview questions 1-3 Tasks and Post-task reflections
		To determine if a sample is unusual, one should examine the sampling distribution of the given statistic, for samples of size n, under the assumption that the null hypothesis describes the population.	CAOS Test of Var 4 Pre-and Post-Open-ended questions 4, 5 Pre-interview questions 2,3,5,7 Post-interview questions 1-3 Tasks and Post-task reflections
Q1	KDU: Variability	Samples are expected to vary.	CAOS Test of Var 6 Pre-interview questions 1, 2, 5, 7 Post-interview questions 1-3 Tasks and Post-task reflections
		The variability of the sampling distribution is influenced by the size of the sample.	CAOS Test of Var 9, 12 CAOS Test of Sig 9 Pre-interview questions 1, 2, 5, 7 Post-interview questions 1-3 Tasks and Post-task reflections
		There is a difference between the variability of	CAOS Test of Var 11

Research Question	Theoretical Framework KDUs	Analytical Framework Understanding Assessed	Data Sources
		the population, the sample, and the sampling distribution.	Pre-interview questions 1, 2, 5, 7 Post-interview questions 1-3 Tasks and Post-task reflections
RQ2: How do simulation tasks influence high school statistics teachers' understanding of simulations and how do they make connections between traditional and simulation approaches for hypothesis testing?	KDU: Simulation Model	1. Establish population parameters 2. Generate samples through simulation 3. Create sampling distribution and assess unusualness	Task data Post-task reflections Post-interview questions
RQ2	KDU: Connecting Approaches	**Step1:** Simulation: What if scenario? Determine a model. Hypothesis test: Statement of null and alternative hypothesis **Step2:** Simulation: Repeat the random sampling or assignment through simulation. Hypothesis test: Check conditions **Step3:** Simulation: Select the appropriate summary statistic Hypothesis Test: Calculate *z* or *t* test statistic	Post-task reflections Post-interview questions

Research Question	Theoretical Framework KDUs	Analytical Framework Understanding Assessed	Data Sources
		Step4: Simulation: Compile summary statistic for distribution formed by simulation Hypothesis Test: Graph the theoretical sampling distribution based on a function **Step5:** Simulation: Assess rareness by finding observed data on simulated sampling distribution and calculating approximate *p*-value. Hypothesis Test: Assess rareness by large or small *p*-value OR large or small test statistic.	

Notice that the first part of the table corresponds to research question one, "How does engaging in simulation tasks for hypothesis testing influence high school statistics teachers' understanding of traditional hypothesis testing?" To answer this question, it was important for me to assess the changes in understanding by comparing the pre- and post-data. Starting with the pre-data, I went through each KDU listed in the table, located the relevant data pieces, and coded it appropriately. For example, the first KDU was logic of hypothesis testing. The corresponding understandings are indirect reasoning will be employed and, therefore, two competing hypotheses are needed; writing the hypothesis to indicate a one- or two-tailed test will address the practical needs of the researcher; hypothesis testing provides a means of "answering" a research question about a population from a sample; failing to reject the null hypothesis does not prove the null hypothesis; and statistical significance does not mean practical significance. The listed

data sources gave me a place to start looking for each understanding. However, I rigorously read all data pieces in the pre-data section to search for additional places where the understanding may have been expressed. This led to restructuring my code list based on a shortened version of each understanding. For example, I realized that I had coded pieces of data as *hypotheses*, which expressed different elements of hypotheses, based on the analytical framework. Therefore, I split my code of hypothesis into the two codes of *indirect reasoning* and *writing hypothesis*. Additionally, I had missed pieces of data that corresponded to understanding that failing to reject the null hypothesis does not prove the null hypothesis. Therefore, I added the code of *proof*. However, from stage one of the analysis process, I had already produced some codes which corresponded to the listed understanding. For example, I had already produced the codes of *alpha, p-value, errors,* and *expected results,* which had encompassed the meaning of a listed understanding. By going through the table, I was able to make sure that I had a code that represented each listed understanding from the analytical framework. The following table (see Table 8) shows the list of codes that were used to represent each understanding.

Table 8

Codes for Analytical Framework for Assessing Understanding of Statistical Hypothesis Testing and Simulations Modified from Smith (2008) and Lane-Getaz and Zieffler (2006).

KDU	Code	Understanding Assessed
Logic of hypothesis testing	Indirect Reasoning	Indirect reasoning will be employed and, therefore, two competing hypotheses are needed.
	Writing hypotheses	Writing the hypothesis to indicate a one- or two-tailed test will address the practical needs of the researcher.
	Infer from sample to population	Hypothesis testing provides a means of "answering" a research question about a population from a sample.
	Proof	Failing to reject the null hypothesis does not prove the null hypothesis. Statistical significance does not mean practical significance.
Probabilistic Nature	Alpha	A "cut point" is necessary to determine whether to reject the null hypothesis or not. This decision considers the probability associated with the sample.
	Errors	A "cut point" determines the probability of a Type I error.
	P-value	The *p*-value is the probability of getting values as extreme or more extreme as the observed value, if the null was true. If the *p*-value is small, this means the result was unlikely if the null was true and the null should be rejected. You have evidence for the alternative. If not, you do not have evidence for the alternative.
Data Collection	Bias	Samples must be unbiased and random.
	Size of sample	Larger samples are more representative of the population.
	Type of Inference	The way in which a sample is chosen will affect the nature of the inference that can be drawn, including the population to which the inference can be applied.
Variability and Sampling Distribution	Variability	There is a difference between the variability of the population, the sample, and the sampling distribution. Samples are expected to vary.

KDU	Code	Understanding Assessed
	Sampling Distribution	For a given sample size, n, and sample statistic, the sampling distribution of the statistic gives a probability distribution of values taken by the sample statistic for all possible samples of size n. It is not the distribution of a particular sample. The variability of the sampling distribution is influenced by the size of the sample.
	Expected results	To determine if a sample is unusual, one should examine the sampling distribution of the given statistic, for samples of size n, under the assumption that the null hypothesis describes the population.
Simulations	Simulations	The steps to conduct a simulation are: 1. Establish population parameters 2. Generate samples through simulation 3. Create sampling distribution and assess unusualness
	Connecting Approaches	Step1: Simulation: What if scenario? Determine a model. Hypothesis test: Statement of null and alternative hypothesis Step2: Simulation: Repeat the random sampling or assignment through simulation. Hypothesis test: Check conditions Step3: Simulation: Select the appropriate summary statistic Hypothesis Test: Calculate *z* or *t* test statistic Step4: Simulation: Compile summary statistic for distribution formed by simulation Hypothesis Test: Graph the theoretical sampling distribution based on a function Step5: Simulation: Assess rareness by finding observed data on simulated sampling distribution and calculating approximate *p*-value. Hypothesis Test: Assess rareness by large or small *p*-value OR large or small test statistic.

These codes were used to classify the data according to the understanding to be assessed. Each time I identified a corresponding data piece, I determined if it provided evidence for that category by comparing it to the listed understanding from the analytical

framework and then wrote about how much of an alignment that data piece showed to the understanding. For example, for the understanding, "If the *p*-value is small, this means the result was unlikely if the null was true and the null should be rejected. You have evidence for the alternative. If not, you do not have evidence for the alternative," I checked if the listed questions from the pre-CAOS test were answered correctly. Then I also checked if the correct conclusion was made on the hypothesis test questions worked on the pre-open-ended response section. Finally, I looked at the responses in the pre-interview concerning *p*-value and determined if the participant had explained the idea expressed by the listed understanding, making analytical memos for each of my assessments. Next, to make an overall conclusion regarding each participants' understanding for each of the KDUs, I reflected and reread the analytical memos for each understanding listed for the KDU. To assist my reflection process, I constructed a descriptive narrative for each of the KDUs, which was used to help report my results in Chapter Four. The process of writing these narratives was an in-depth reflection, which served to provide a synthesis of the data concerning my participants' overall CCK for hypothesis testing. I analyzed all five KDUs for the pre-data section in this manner and wrote a descriptive narrative for each. The KDUs from this study's theoretical framework offered a natural way to organize my narratives based on categories of CCK developed from the literature review in Chapter Two. This organizational scheme was important because it provided a way to compare changes in understanding from the pre- and post-data by categories. After completing the pre-data section, I went through the post-data similarly, making notes and writing analytical memos by summarizing my evidence for each KDU. The synthesis process for the post-data was comparable to the pre-data in that

I created descriptive narratives through reflection and rereading of my summaries of evidence for each understanding. However, I also compared the analytical memos for each KDU of the post-data to the pre-data narratives to check for changes in understanding and included this analysis as part of the post-data narratives, which were also organized based on the five KDUs for CCK listed in the analytical framework. This served to answer research question one by providing an analysis of the participants' CCK before and after the simulation tasks.

The last two KDUs listed on the analytical framework were aligned with research question two, "How do simulation tasks influence high school statistics teachers' understanding of simulations and how do they make connections between traditional and simulation approaches for hypothesis testing?" These KDUs were simulation model and connecting approaches. The main pieces of data were in the post-task reflections that revealed how the participants understood simulations and connected the approaches. For each of the post-task reflections, I created tables comparing the participants' steps of a simulation to a modified version of Lane-Getaz and Zieffler's (2006) SPM, which is listed in the analytical framework for the SCK of simulations, and I also made notes concerning the similarities and differences. This provided me with a way to see how my participants' understanding of simulations compared with the literature. Additionally, I used a modified version of Lane-Getaz and Zieffler's (2006) model for connecting approaches, listed in the analytical framework for the SCK of connecting approaches, and compared this with how the participants expressed how the approaches were connected. I compared their model with the participants by constructing comparison tables and writing about the similarities and differences that I found. To synthesize the results of my

analysis, I created descriptive narratives for the post-task data sections. I created these by reflecting on all data pieces, tables, and notes that were produced from my analysis. These descriptive narratives answered research question two by showing how the participants understood simulations and how they connected approaches.

After I completed this process, I went back to my code list and reorganized them based on how I had organized my descriptive narratives, which was a result of the categories from the analytical framework. The following table (see Table 9) shows how my final list of codes was grouped by the five KDUs of CCK of hypothesis testing and the two KDUs of SCK from the analytical framework. However, what is important to note is that a list of codes did not fit into these groups. These codes were produced inductively during stage one of data analysis. I named this group influencing factors. This final group of codes would be used to produce my overall themes as reported in phase five of my data analysis steps.

Table 9

Final Code List and Categories

Group	Codes
Logic of hypothesis testing	Indirect Reasoning Writing hypotheses Infer from sample to population Proof
Probabilistic Nature	Alpha *P*-value Errors
Data Collection	Bias Size of sample Type of Inference
Variability and Sampling Distribution	Variability Sampling Distribution Expected results
Simulations	Simulations Connecting Approaches
Influencing Factors	Visualization Procedures Lesson Plan Context

Stage three involved going back through my data and checking that I had not missed any pieces of data that fell into one of these categories. Additionally, I reread all narratives to ensure that the analysis was consistent with the data. At this point, I had answered research question one by showing the changes in understanding when comparing the pre-data with the post-data. Also, narratives produced from the post-task reflection data showed how the participants understood simulations and how they connected approaches. However, I also wanted to dive deeper into the data and produce overall themes for each research question, which addressed why these changes in understanding occurred and what influenced how the participants understood simulations

and connected approaches. This was done in stage four of my analysis, developing themes.

The process of developing themes, which is a "formidable and daunting task," can only be "accomplished through deep reflection" (Saldana, 2016, p. 281). Although this step is listed separately, this process was ongoing throughout the entire process. It began with the inductive codes that were produced and classified as influencing factors (see Table 9). I had named this group of codes influencing factors because they referenced pieces of the data which illuminated more of why changes were happening or what was influencing the participants' understanding instead of assessing what the understanding was. These codes were *visualization, procedures, lesson plan,* and *context*. However, although I had initially included context as an influencing factor, after completing this piece of the analysis, which involved serious reflection and comparing results to the data, I determined that context was important to the participants as statistics educators in terms of interpreting and making sense of results, but the data pieces corresponding to this code did not highlight an instance of influencing the participants' understanding. Therefore, the three codes of *visualization, procedures,* and *lesson plan* were the most important pieces in determining the overall themes.

After rereading the data pieces corresponding to the influencing factors category, I still had not determined the overall themes for each research question. Therefore, I decided to construct additional narratives corresponding to the task phases to serve as a reflection tool. These narratives are reported at the beginning of Chapter Four. After constructing these new narratives and rereading the other narratives, I noticed that the participants focused on concepts, specifically on the indirect reasoning of hypothesis

testing. With this new perspective of looking for my participants focus on concepts and by considering the data pieces in the influencing categories group, I produced the overall themes for each research question, which will be reported in Chapter Four.

After completing this process, I began the fifth stage of analyzing the second case. I repeated steps two through four, like I did with the first case. I used the analytical framework to organize my data by categories, assessed understandings, and wrote descriptive narratives for the pre-data, post-data, and post-simulation task data. Then I reflected on this second case by rereading the data, memos, and narratives, to determine the themes. Because I produced different themes for my second case, I went back over the data pieces for both cases to check the accuracies of my themes. Based on this comparison and rechecking the data, I modified the themes for my participants. For example, for Kathleen, I had produced a theme regarding visualization. By having this new lens to look through when rereading the data pieces and descriptive narratives, I realized that Carrie's data provided evidence for this theme as well. Also, my original theme for Kathleen was that the discussion component of the tasks was what influenced her understanding of the KDU logic of hypothesis testing. However, for Carrie what influenced her understanding of this KDU was the focus of the simulation tasks on concepts instead of procedures. By rereading my evidence for Kathleen, I determined that the discussion aspect that I was writing about was the concepts. I had not made the final, logical step of considering what was being discussed until I compared the themes and rechecked the evidence. Through this process of comparing themes, reflecting, and rereading my evidence I was able to revise my themes and produce results rooted and supported by the data.

Finally, I completed stage six, a cross-case analysis. I first compared the initial understanding of hypothesis testing of each participant, along with as assessment of the changes in understanding for each participant as shown in the post-data. To do so, I constructed tables to analyze these similarities and differences. Next, I compared how each participant understood simulations and how they connected approaches. For each comparison, I used tables to relate the corresponding elements and discussed the similarities and differences. Finally, I compared the overall themes for each case. Although I had produced similar themes, I produced a matrix to compare the evidence of each theme for the participants. In this manner, I discovered similarities and differences between the two cases as reported in Chapter Four. Next, I will discuss how the quality and credibility of this analysis and the rest of this study was achieved.

Quality and Credibility

Quality and credibility was ensured in several ways. First, I achieved triangulation through the use of several data sources (Yin, 2014). Basic teacher knowledge was identified with the CAOS assessment items. I was then able to compare these results with responses in the tasks and open-ended questions portion. Finally, to confirm my results through triangulation, I asked additional questions during the interview portions. I also strengthened the results of this study by using multiple cases to identify similarities and differences (Yin, 2014). I also sought alternative conclusions, disconfirming evidence, and used logic to determine the explanations that best fit the evidence. I did so by thinking about why participants may have answered certain questions correctly or incorrectly on the CAOS test and used evidence from the open-ended questions and task responses to confirm or disconfirm my theories regarding the participants' understanding.

If evidence of thinking was not apparent in written responses, I asked questions in the interview portion to elicit the participants' understanding. I also used an iterative process by revisiting all relevant portions of the data several times to check for accuracy.

Limitations and Delimitations

Limitations are factors that impact the study beyond the control of the researcher, and delimitations are influencing factors under the control of the researcher. In this section, I will discuss both types of influences, along with rationales behind the decisions that I made. I chose mathematics teachers of AP and equivalent non-AP statistics classes, instead of general mathematics teachers who teach some statistics, because inferential topics are more heavily covered in these types of classes. Additionally, I did not choose teachers who had no knowledge of hypothesis testing, because I wanted them to be able to articulate their thinking about this topic. I also decided to have participants work the tasks as a group, instead of individually, because the tasks were designed to include a discussion element.

Another decision that I made was regarding the types of tasks that I used. I selected tasks that were created using a six-phase lesson structure based on NCTM's mathematical teaching practices. Because these are research-based practices, I believed this was the best choice for influencing the participants' knowledge.

I wanted to choose teachers who had not used simulations for inference previously. However, as already explained in this chapter, Chase did not realize that he had been exposed to using simulations for inference in his AP training. Therefore, his knowledge may have impacted the implementation of the tasks. One goal of the study was to provide information regarding how these tasks can be used in professional

development settings to increase teachers' understanding of hypothesis testing. Therefore, this affects the transferability of the results slightly, because typically teachers attending professional development using these types of tasks would not have seen simulations for hypothesis testing previously.

Chapter Summary

I selected an explanatory multiple-case study design to investigate the changes that occurred in high school statistics teachers' understanding of hypothesis testing when engaged in simulation tasks. Additionally, I investigated how they understood simulations and how they connected simulation and traditional approaches. I analyzed my data using a rigorous multiple step approach rooted in the data. Additionally, I conducted a cross-case analysis of my themes to strengthen the results of this study. I will share the results of this study in the next chapter.

CHAPTER FOUR: RESULTS

The purpose of this study was to answer the following two research questions:

1. How does engaging in simulation tasks for hypothesis testing influence high school statistics teachers' understanding of traditional hypothesis testing?
2. How do simulation tasks influence high school statistics teachers' understanding of simulations and how do they make connections between traditional and simulation approaches for hypothesis testing?

Recall from Chapter Two that CCK refers to basic knowledge of the subject matter that is not specific to teaching (Ball et al., 2008). Therefore, research question one focused on the changes seen in the participant's CCK after engaging in the simulation tasks. Also, the content knowledge that is unique to teaching is referred to as SCK. This knowledge is typically only needed for teachers. Based on the premise that teachers can use simulations in the classroom to help students foster a deeper understanding of hypothesis testing, understanding simulations and how simulation approaches are connected to traditional approaches, which research question two investigated, would fall under the category of SCK. An explanatory multiple case study was used to answer the research questions of interest. In this chapter I will present the results from the data analysis I conducted, as described in Chapter 3. I will begin with a brief introduction of each participant. Also, because the simulation tasks were integral in changing the participants' content knowledge and influenced how they understood simulations, I will provide a detailed narrative of how the participants engaged in each task. Then, I will answer research

question one and two for each case. Finally, I will complete the chapter with a cross-case analysis.

Participants

I selected three high school mathematics teachers as my participants. The following table (see Table 10) provides background information for each.

Table 10

Background information of participants

Name	Degree(s)	Training to teach statistics	# years teaching	# years teaching statistics
Carrie	BS Mathematics M.Ed. Math Education	None	8	6
Kathleen	BS Mathematics MS Education	Statewide dual credit training	21	8
Chase	BS Mathematics MEd Teaching and Curriculum EdS Curriculum and Instruction	AP summer institute. Statistical thinking common core training	8	0

The first participant, Carrie, is a 31-year old high school teacher with eight years of teaching experience in mathematics, including six of those years also teaching non-AP statistics. She holds a master's degree in mathematics education and worked through lessons in her statistics book to teach herself statistics. Carrie was extremely excited to be a part of this study. She felt that she did not possess a conceptual understanding of

statistics, because she had never been taught this subject before. Also, she began teaching statistics because no one else wanted to teach it.

The next participant, Kathleen, is a 57-year-old woman who has taught mathematics for 21 years and statistics for eight of those years. She has a BS in Mathematics and a MS in education. She received statewide training to teach dual credit statistics but felt that she mainly learned statistics on her own. The statewide dual credit class is like an AP statistics course in that students must pass a test at the end of the year to receive college credit. In her pre-interview, when she was talking about her college class in statistics, she said, "I think I spent two weeks not having a clue what he was talking about, because I had never seen it before. I didn't have any idea" (October 19th, 2017). Kathleen also felt that she still had many topics in statistics that she still needed to learn better, especially hypothesis testing.

The final participant, Chase, is a 32-year old male who had been teaching for eight years. He had tutored statistics and received AP training, but he had never taught statistics before. I was surprised that the pre-data revealed that Chase seemed to possess the strongest content knowledge of hypothesis testing, because he had the least experience with the subject. However, in question five on the post-open-ended response survey, which asked if there was anything about hypothesis testing that the tasks made you think about differently, Chase wrote, "No, but probably only because I have done simulations before (undergrad coursework and at AP Stats training)." (November 27th, 2018). It was when I read this answer that I discovered that Chase was not representative of my population of interest, because he had engaged in simulations for hypothesis testing prior to the study. Chase had mistakenly believed that he met the requirements for

study, because I had used different terminology than the AP trainer, who referred to this method as just simulations instead of using simulations for hypothesis testing like I did. Therefore, his data will not be shared in this chapter, but information regarding Chase is included here because of the importance of some of his dialogue with the other participants during the tasks. These tasks were the intervention used in this study; therefore, it is important for the reader to be provided with the details regarding how the participants interacted with the tasks, which will be shared next.

Simulation Tasks

Each simulation task lesson plan was described in detail in Chapter Three. However, I will provide a brief overview of the lesson plan template used to create each task before describing what happened with the participants in this study. The lessons were designed based on a six-phase lesson structure, which aligned with teacher actions advocated by NCTM's (2014) *Principles to Actions: Ensuring Mathematical Success for All* (see Strayer & Matuszewski, 2016). The six phases are (1) commitment to a position in a rich context, (2) statement of possible hypotheses, (3) statement of expected results assuming the null hypothesis is true, (4) revelation of study results, (5) simulation under the null hypothesis, and (6) making a conclusion. In phase one, by having students commit to a position in a rich context, students become interested and engaged in the lesson. The second phase has students discuss all the possible hypotheses that could be true for the given scenario and to list them. Next, the null hypothesis is identified, and students are asked to determine the most likely outcome if the null hypothesis is true. However, their thinking is pushed to acknowledge variability by also listing the types of results that would not be surprising if the null hypothesis were true. Student interest is

continued in the next phase by revealing the actual results from the study. For phase five, instead of providing steps to a pre-determined simulation method, students create their own simulation of the null hypothesis scenario to better understand the purpose of the simulation. After performing several simulations on their own, technology is used to construct an empirical sampling distribution and obtain an accurate estimation of the *p*-value. Finally, students use the simulation results from technology to make a conclusion based on the sample data. This design was used for each of the tasks, and I will describe how the participants engaged in all three simulation tasks next.

Task A

Helper or Hinderer was selected as the first task because of its simplicity. It involved only one sample with a 50-50 probability and used counts instead of a mean or proportion. Therefore, this task could be considered one of the easier simulations to design. The following shows the scenario used for task A.

> In the original study, conducted by Yale researchers in 2007, groups of 6-month-olds and 10-month-olds watched a puppet show with neutral wooden figures, where one figure, the climber, was trying to get up a hill. In one scenario, one of the other figures, called the helper, assisted the climber up the hill. In the other scenario, a third figure, called the hinderer, pushed the climber down (Holcomb et al., 2010).

After watching a video of a puppet show, which I showed the participants to illustrate the helper and hinderer scenario, the participants were asked, “If 16 pre-verbal children participated in this study, how many do you think chose the helper toy? Why? What factors do you think might be at play when the children make their choice?” Carrie’s

initial hypothesis was that the children would not have a preference and stated that eight of the sixteen would select the helper toy. However, Kathleen believed that 12 would select the helper toy because children had a sense of social justice. Chase selected 10 and wrote, “some preverbal kids have a sense of right/wrong but most may be choosing at random”.

After each participant committed to their own hypothesis for the scenario in question one, the second question asked them to list all the different possible hypotheses for the scenario. The group decided that the experiment was about determining if children have a sense of social justice or not at that age, which could be determined by whether they selected the helper toy. Therefore, they identified two competing hypotheses: no preference and preferring the helper toy. They also mentioned that it was possible that the children could prefer the hinderer toy. Carrie joked, “I mean I have a child that prefers villains”. Both Kathleen and Chase wrote the hypotheses out in words, but Carrie used typical hypothesis test notation of $p = .5$ and $p > .5$. She also used the symbol, p, for proportions, even though the question was asked in terms of counts.

Next, question three had the participants discuss what the most likely outcome (for number of infants choosing the helper toy) would be if the study was conducted with 16 infants and the null hypothesis of no preference was true. They were also asked to list what other results would not surprise them. Kathleen and Chase provided a very narrow range of values of eight to nine on their handout, with Chase adding in parentheses maybe 10. Kathleen and Chase were both unsure about how much variability should be expected, as shown in the following excerpt.

Kathleen: See that's what I was thinking about, what Chase said. Like if it was just 10, I might be debating that. Like was it just a fluke, but then when you get to a number like 12, then you think that like, "Oh definitely. Yes. They definitely have a sense of social justice."

Chase: So, the cut off for you is like 10 or 11? Well, what about 11? What would you do with 11?

Kathleen: I don't know. What I would do?

However, Carrie was more accurate in predicting what types of counts would not be surprising before conducting the simulation, with her answer on the handout of a range of 5-11 infants selecting the helper toy would not be surprising if there was no preference. She did not state her reasoning regarding this question in the discussion though.

Next, I revealed that in the actual study 14 out of 16 infants chose the helper toy. In response to the question on the task handout, "Do you find the researchers results surprising", Kathleen wrote, "Yes! That's a large majority and tough to comprehend if there is truly no preference". Both Chase and Carrie also stated that it was surprising. Carrie wrote, "that is well outside my range of acceptability", and Chase echoed this sentiment by writing, "this is definitely outside the expected zone if we assume no preference." To determine if their predictions concerning what types of results would not be surprising if the null hypothesis were true were accurate, the participants were asked to design a simulation which corresponded to the situation. The following shows their discussion concerning how to design the simulation.

Carrie: So, based on 50-50 probability.

Kathleen: So, like roll a die and let even could be the helper and odd …

Chase: We can use cards like red and black, but I don't trust the shuffling.

Carrie: Yeah, I don't like the cards. We could use a coin.

Researcher: So, you're saying flip a coin or toss a die. What does that mean? How many times would you flip it for one trial?

Carrie: 16.

Researcher: And record what?

Chase: Number of heads.

Researcher: What did you let heads represent?

Chase: The helper toy.

When designing the simulation, Carrie correctly identified that we would be simulating the null hypothesis, which meant they needed to ensure that each child had a 50% chance of selecting each toy. Additionally, she stated that a single trial meant to flip a coin for a total of 16 times. Kathleen also had no trouble designing a simulation and quickly suggested rolling a die with an even number representing picking the helper toy and odd representing the hinderer toy. They decided to flip a coin for the simulation, and I had them conduct several trials on their own. However, with only three people conducting the simulations there was not enough data to construct a meaningful dot plot. Therefore, I showed them a sample of a typical dot plot that could have been obtained in a classroom setting. For this dot plot, 30 trials were conducted, but none of the trials resulted in a number as large as 14. Therefore, all the participants agreed that the researcher's results were surprising. However, when I asked them if 30 was enough trials, they all said no. At this point, I showed them a free online software program, *StatKey* (http://www.lock5stat.com/StatKey/), which would construct an empirical sampling

distribution. I modeled how to change the settings to perform simulations based on the null hypothesis of the infants randomly selecting a toy being true. I initially conducted one trial at a time to emphasize that the technology was doing the same thing as their hands-on simulation with the coin. Then, I clicked on produce 1000 trials and had the software highlight in red the number of trials which resulted in a count of obtaining 14 or more infants choosing the helper toy if there were no preference. The technology also produced an estimated *p*-value, by calculating the proportion of times that 14 or more occurred. However, I did not explicitly refer to this number as a *p*-value.

After using the technology to produce the simulated sampling distribution and the estimated *p*-value, the participants were asked, "Based on this simulation how surprising are the actual results of the study? In other words, how likely would it be to obtain 14 or more out of the 16 infants choosing the helper toy," Carrie wrote, "Very surprising. Not likely that 14/16 would randomly choose the toy" However, both Kathleen and Chase used the estimated *p*-value in their answers. Kathleen wrote, "Very surprising. There is less than ½ percent chance that 14 out of 16 would choose the helper if there is no preference", and Chase wrote, "According to the simulation it is surprising that the researchers found 14 who chose the helper since our simulation only showed the likelihood of 14 being chosen as less than ½ percent chance".

After asking how surprising the results were, the participants were then asked, "Based on the simulation, what conclusion should the researcher draw?" They were all able to make an appropriate conclusion. At the beginning of the task they stated that if more children choose the helper toy then that means they have a sense of social justice, and they wrote their conclusions on their task handout in terms of this context. Kathleen

wrote, “They should conclude that the infants have a sense of social justice”, and Carrie wrote, “There is a sense of social justice and right vs. wrong at play in order for so many to choose the helper”.

Overall, Carrie contributed the least to the discussion during the task. For example, Chase and Kathleen discussed the question about the different hypotheses and what results would be typical, but Carrie just wrote her answers on the handout. Also, both Kathleen and Chase focused more on the context of the problem. For example, they wrote the hypotheses in terms of social justice, and Carrie used hypothesis test notation involving proportions. Carrie did contribute more to the discussion on designing the simulation and seemed to be more confident during that part of the activity. After participating in this initial task, the participants became familiar with the overall lesson plan design and more quickly answered questions in the following two tasks. The second task will be described next.

Task B

The second simulation task involved comparing average test scores on yellow and white paper, as shown in the provided scenario below.

> Math teachers often use two different forms of an exam to prevent students from cheating. One trick teachers use is to give the same exam but on two different colors of paper (white and yellow). Some students believe that yellow is a happier, peaceful color compared to the stark white and that they would tend to score better on yellow paper. To investigate this claim, a teacher gave all of her students the same test and randomly chose half the students to take it on white paper and half the students to take it on yellow paper (Statcrunch, n.d.).

Question one asked, “If 20 students took an exam, 10 on white paper and 10 on yellow paper, how do you think the average score from students who took the exam on yellow paper would compare to average score of the students who took the exam on white paper? Why?” Kathleen and Chase believed that the scores would be higher on the yellow paper. However, Carrie disagreed and wrote, “I think the average scores will be close to the same. I don’t think the color matters.”

For the second question, which asked them to list all the possible hypotheses, Kathleen answered first and said, “They could be the same thing, or yellow could be higher, or yellow could be lower.” She also started to add not equal to her list. Chase noted that not equal to would be the same as grouping less than and greater than together. So, Kathleen left her original three hypotheses.

Next, question three focused on what would be expected if the null hypothesis was true by asking, “In statistics, we typically subtract the average scores from two groups in order to compare them. If the color of the exam did not affect students’ scores, what would be the most likely outcome (difference in the average scores) when this study is conducted with 20 participants?” They all quickly agreed for this question that it would be zero but debated the next question, which asked them to list what types of results would not surprise them if the null hypothesis was true.

Kathleen: I mean I would think that it could vary two points maybe either way.

Chase: I would probably give it more than that.

Carrie: I wouldn’t be surprised if it was five points off. If it was more, I would be surprised by that.

Kathleen: And I kinda look at that in my own classes too. You know if second period's average was an 80 and really if another class averages a 75, I'm not shocked about that.

Chase: That's why I was going to give it more of a range of 10.

Kathleen: I don't know about 10.

Carrie: You expect some variability because it is different students.

Chase: But are you thinking 5 away from the mean each way?

Kathleen: Correct. Yes. Yes.

Carrie: So negative five to five.

Researcher: So, Chase, were you sticking with five or are you thinking ten?

Chase: No, I was thinking like the range of value would be 10.

Researcher: Oh, so your saying negative five to five.

Chase: Yeah.

At first, Kathleen was very conservative with her range of only two. However, when the participants starting to think of their own experience and apply the real-world context of comparing their own classes' test scores, they expanded their guesses to negative five to five, which was still a low range for the expected variability as will be seen in the simulation. Also, Carrie noted that some variability should arise naturally and would be expected because different students took the test.

Next, I revealed that the actual difference was 6.3 points, which based on the participants' previous responses would be surprising. The participants then conducted a hands-on simulation using slips of paper with the original 20 test scores. The participants decided to shuffle the scores and divide the cards into two piles. The first 10 selected

would represent scores on yellow paper and the remaining would be white paper. They averaged each group of scores and subtracted the yellow average minus the white average. After conducting a few trials themselves to understand the simulation, we turned to technology to produce many trials. When the simulated sampling distribution was constructed using technology, Carrie wrote that she "expected less variability." (Task B Handout, November 15th, 2017). She thought that the difference of 6.3 from the study would be large enough to reject the null hypothesis, but the simulated *p*-value was .127. This number resulted in an interesting discussion.

Researcher: What are we thinking? So, is it surprising?

Chase: It's not as clear as the last one (referring to Task A). Like this one is more likely to happen.

Kathleen: So, there's a chance that that could happen.

Chase: So, there is a 13 percent chance that this could just happen randomly.

[several moments of silence with the participants looking at the distribution]

Researcher: So, these are the results we would expect if the true mean was zero. So, what do you [Chase jumps in quickly with a response].

Chase: If I were a student it would be compelling to me, and I would want yellow paper every time. Because I mean 13 percent is not that big. Right? But would I go publish a paper in like an academic journal or something?

Researcher: So that is where I start having a conversation with students like you know, really what is surprising and compelling evidence? The

five percent seems kind of arbitrary. And you know we can kind of go back to the jury thing. Well, I mean, is this enough to say someone is guilty? Is this enough evidence that the yellow paper really is making you perform better on your test?

Kathleen: No.

Researcher: You know some people do use point one. Right? And this is close to, close that. So, what? I don't know. What conclusion would you draw?

Kathleen: I would still, based on that number, I would still have to say as much as I would like to give them colored paper, and I think that might calm them a little. It's not enough to convince me that the colored paper is better.

Researcher: So, kind of based on what you both are saying is like there's some evidence. There's some. It's making you a little suspicious, but not necessarily enough to conclude that the yellow is really causing it?

Chase: I would maybe even talk to my kids, if I were teaching this. I would talk to my kids about like the difference between hard skills in an industry and soft skills in an industry. And if you are like a data analyst, and you're seeing this. You have to then present these findings to someone, and you have to have the hard skills to know what it means like legit. But then you have to be able to have the skills to know the population you're talking to and is this going to be important for them? Is that going to be compelling for them or

not? Because if you're talking to a group of students, and they get a free choice. There's no harm in picking the yellow paper right? So, you would maybe say, yeah that's good enough evidence for you. But if you're, you know in, if you're like in CSI, like you probably have a higher standard than 13 percent. Like that's probably not good enough.

Kathleen: And that yellow paper is more expensive.

At this time, I was going to ask more about the alpha-level, but Carrie interrupted with a question about the validity of the simulation approach.

Carrie: Ok so here's my question. Don't we need to be looking at the sampling distribution.

Researcher: So, what is this? (pointing to the simulated sampling distribution).

Carrie: I mean this is just a simulation. This is just one thousand six hundred and eighty trials of this experiment. When we do the calculations the old way, that's not just simulations that the sampling distribution, that's all possible samples.

Researcher: That's the theoretical sampling distribution.

Carrie: Yes. So how much of a difference would that make? I guess my question is if you are to keep generating samples, like double the amount that you have now, how much would that percentage change?

Chase: So theoretically the mean would go to zero.

Carrie: I mean yeah, the mean will, but if you were to keep generating samples, I'm talking about the percentage (referencing the simulated *p*-value), how much change are we going to see?

Researcher: So, what was it before, .127? Ok, so I am going to generate another 1000.

Chase: It will normalize and go to zero.

Carrie: I know that, and I'm asking how much that percentage could change? Like if we did it the old way. I know we could use the normal curve and calculate the test statistic.

Chase: That's what this is doing too. That number (talking about the *p*-value) is going to get, it's going to approach the number that would be on the normal curve.

Carrie: That's what I thought. That's all I was asking was that. How reliable we were, or do we need the full-on sampling distribution?

Researcher: So, I just did this almost 4000 times (referring to the simulated sampling distribution).

Chase: And I think the answer is just it's the more samples you do the more reliable it is.

Researcher: Look at the difference, right? So, we were at .127 and now we are at .125.

Chase: But look where the mean is now.

Kathleen: Yeah, it's, that's the mean. Yeah. Right. (They are looking at how the mean of the simulated sampling distribution was .0001 and is now 0, which they had stated earlier that it should be).

Chase: And your kids, if any of your kids have had calculus, and if they understand the concept of a limit, they're going to get that like the simulations are, if you kept doing them over and over again, you're approaching that limit, which is what the textbook is like.

Upon reflection, this point in the conversation could have been a good opportunity to work this problem out with the traditional method and compare. However, I was still thinking about the previous conversation about the alpha level and steered the conversation back to this topic. I changed the area highlighted by technology to shade an area of .05 to the right instead of shading the *p*-value area. The difference, which corresponded to the cutoff of .05, was 8.9 points.

Researcher: That .05 we were talking about. So, what does this mean?

Chase: So that 8.9 is a difference that we would expect to see between the means if there was a significant amount.

Kathleen: So almost nine points higher on the yellow paper.

Researcher: So, if it was nine points higher, then we would say what?

Kathleen: That there's a difference.

Researcher: So, could we get like a number of 10 though if there is actually no difference?

Carrie: Yes, but that would be really rare that happened by chance.

Researcher: So, it's rare to happen, but it could happen?

Chase: A false positive.

Researcher: So, let's make it clear, we would conclude that there is a difference but there wouldn't actually be one. What's that called?

Carrie: I have to make a little chart.

Kathleen: So, we reject when we shouldn't have, so a Type I.

Researcher: So what percent of the time will we do that if the null is true?

Chase: Five percent.

The pre-data had indicated that Kathleen did not understand the relationship between the alpha-level and the Type I error. Kathleen knew the name for the type of error, and Chase figured out that the probability of this happening would be 5 percent. As will be indicated in the post-data, Kathleen would now be able to explain this relationship.

Finally, to end the activity, the participants wrote on their handout what conclusion the researchers should draw if their data showed a difference of 6.3. Kathleen wrote, "There is no difference in the color of the paper when comparing test scores." Carrie wrote, "This could have just happened by chance." Next, I will share how the participants engaged in the last task.

Task C

The dolphin simulation task corresponded to a two-sample *z*-test for proportions. Out of the three tasks, designing the simulation is the most difficult for this scenario. I selected this task last so that the participants would have had some experience with designing simulations previously. Here is the scenario.

Swimming with dolphins can certainly be fun, but is it also therapeutic for patients suffering from clinical depression? To investigate this possibility,

researchers recruited 30 subjects aged 18-65 with a clinical diagnosis of mild to moderate depression. Subjects were required to discontinue use of any antidepressant drugs or psychotherapy for four weeks prior to the experiment, and throughout the experiment. These 30 subjects went to an island off the coast of Honduras, where they were randomly assigned to one of two treatment groups. Both groups engaged in the same amount of swimming and snorkeling each day, but one group did so in the presence of bottlenose dolphins and the other group did not. At the end of two weeks, each subjects' level of depression was evaluated, as it had been at the beginning of the study, and it was determined whether they showed "substantial improvement" (reducing their level of depression) by the end of the study (Rossman, 2008).

I revealed part of the study's results at the beginning that 13 of the 30 participants showed substantial improvement and asked them how many of the 13 improvers did they think were in the dolphin group.

Chase: I'm going to say 10.

Researcher: OK. Why?

Chase: I feel like it should be more than half of the improvers. So like half would be like six and a half, right? So, say seven. It's like half. So more than that. But I also think that some of the people probably just improved by changing location and being, you know being in a study like having the attention and the structure of being in a study.

Carrie: Going on a vacation. Yeah.

Researcher: So, you would say improvers would be in the non-dolphin group anyway.

Carrie: Yes.

Researcher: Do you think it's as high as Chase?

Kathleen: I probably think so too. I think maybe 9 or 10 at least.

Carrie: I agree with that.

Researcher: So, you're thinking that maybe the dolphins help?

Kathleen: Yeah.

Researcher: So, you all said 9 or 10?

Kathleen: Yeah, so 11 would not surprise me, but I think maybe 12 or 13 would surprise me.

The first question only asked the participants to state what they believed the outcome would be, but Kathleen was already thinking about what types of results would be surprising, which is not asked until question three. Question two asked them to list the possible hypotheses. The participants hesitated to answer at first. To assist them, I asked them what they thought the researchers were trying to show.

Kathleen: That swimming with dolphins was therapeutic.

Chase: That more improvers would be in the dolphin group.

Researcher: Is that the same thing?

Chase: Yes.

Researcher: So, what else do we have?

Carrie: Does not help.

Researcher: So, if it does not help, what does that say about the number of improvers?

Chase: About the same for the dolphin and control group.

Researcher: So, are there any more?

Carrie: I mean I guess you could say that it makes depression worse. (laughing)

They determined that the hypotheses would be that either the number of improvers would be the same or that more would be in the dolphin group. Although Carrie mentioned that a third hypothesis of swimming with dolphins may make depression worse, none of the participants listed this as a hypothesis. Next, the participants were asked what the most likely outcome (difference in number of improvers between the dolphin and the control group) would be.

Researcher: So, what's the most likely outcome if there is no difference?

Chase: So, one.

Researcher: Why?

Chase: Well because half is six or seven. So, there's at least a difference of one.

Researcher: So, what about zero?

Kathleen: We can't get that at all. We're not going to be able to get zero. Are we?

Researcher: So, what about negative one?

Chase: Yeah.

All the participants agreed that you cannot obtain zero and wrote on their handout that negative one or positive one was the most likely outcome. Next, they were asked to determine what types of results would not surprise them if the null hypothesis were true.

Researcher: Ok, so still assuming the null is true, what kind of results would not surprise you? So negative one to one is most likely. What would not surprise you?

Kathleen: Oh really, even with a difference of two though, it would not surprise.

Chase: You know I think it's even more than that. Yeah, three or negative three.

Carrie: Am I overanalyzing at this point? When I'm looking at all the different numbers, it's not possible to have a difference of two. I am just listing all the possible differences. So, getting a 2 is not really possible. I mean, do you want your students to think about that?

Chase: I mean yeah, because we are talking about in the real world.

Carrie had listed all the possible differences on the side of her paper and correctly noticed that you could only obtain an odd number for a difference in this scenario, because the number of improvers was 13. Kathleen still wrote a difference of two in either direction on her handout, but Chase and Carrie wrote negative three to positive three.

Next, I revealed the actual study's results that 10 of the improvers were in the dolphin group and three were in the control group. Chase was very excited that his prediction from question one ended up being correct. He said, "I'm killing it." They all

believed that this difference of seven was surprising and evidence that swimming with dolphins helped.

For the next question, the participants were asked to design a simulation to represent this study under the assumption that swimming with dolphins does not help. I provided the participants with cards, dice, and coins.

Kathleen: So, we could do the cards, couldn't we?

Carrie: Yeah draw 30, and do black or red.

Kathleen: Black or red. So, black could be the control group.

Chase: So, that's like heads or tails. Like flipping a coin.

Researcher: So, one thing to think about is just how were you representing the participants.

Chase: So, first we need to separate them into two groups.

Kathleen: So, we can do red and black.

Chase: Right.

Researcher: And how are you representing who is going to improve?

Chase: And then after you draw those, it's almost like you need to do another, like flip a coin.

Kathleen: We could make odd cards the non-improvers and even the improvers

Researcher: But how many improvers were there?

Kathleen: 13.

Researcher: So, there's going to be 13 no matter what.

Chase: So, we have to get 13.

Carrie: So, count out 13 red cards, and let's see how many black cards? How many were there?

Kathleen: Don't we need to do the odd and the even and the red and the black? So isn't this like the dolphin and control group (pointing at red and black cards), and then the even cards could be the improvers and the odd cards could be the . . .

Chase: But you're not going to ensure that you get 13.

Kathleen: But I mean we would have to hand set it up. Does that make sense?

Chase: Ok, here's another way we could do it. You could draw out, you could assign them to their groups here. Each person is a card, right, so you could shuffle them all up and just draw out 13 cards and let the first 13 be improvers. The first 13 are improvers and then you can mark which group they were in.

Carrie: That's what I was thinking just like 13 red and 17 black cards. So we have 30, and then shuffle them up and then just take out the first 13.

Researcher: So, I think we're kind of mixing maybe two different ideas here. So, you're saying since you already know 13 are going to improve, you've designated them red cards and the rest of them are black. Because that's going to happen no matter what. So, we know we know the end result. Right?

Chase: So, you would need to take yours and split them in half. Because you've already designated who's improving and who's not. And I

wasn't. I was splitting them in half first and then designating who improved and who hasn't. It doesn't matter.

Kathleen: I was doing both of them.

Researcher: This is kind of tricky. So usually what I try to say is OK decide how are you representing the participants? So, you said the participants are the cards. And then also think about how are you going to determine who's an improver and who's not an improver and are you guaranteed there's 13 of them? And then the other thing is, then how are you making sure that the improvers and not improvers were just randomly assigned to the two groups because it didn't matter which group they were in.

Chase: I mean honestly if I were designing this experiment, I would, I would flip a coin and split them evenly, and then I would put all their names in a hat, and that's how I would do it. They would have names, like the participants would. I would put their names in a hat and then draw from the hat, and just draw 13. Because that's less confusing than representing it with a number or a red.

Kathleen: So then would we start with the improvers like what she's saying.

Chase: It doesn't matter. It can be either way.

Researcher: I want to go back and understand what your method was. (talking to Chase). So, you were saying you put them all in a hat and then pull out like 13. So how would you know who was in treatment and who was the control?

Chase: So, I would have already split them up.

Researcher: Oh, yours is a two-step process then. OK. OK. I missed all.

Carrie: I want to understand everybody's idea. I just, I think what I was trying to say is that you have 13 red and 17 black, and I'm going to turn them over and shuffle them up, and like 15-15. So, like that's the treatment and control. And then look at my treatment group and see how many improvers and look at my control.

Researcher: Yeah, yeah, that's the most common method.

Carrie: I wanted to know if I was right or wrong.

This design is difficult because to simulate the null hypothesis being true you need to consider the results of the study that there were 13 improvers. As Chase noted, you could either assign the improvers before or after the randomization occurs. Carrie did determine a legitimate design and was able to explain her method clearly at the end, although she had originally said to just pick 13 instead of the 15-15 split needed for the treatment and control groups. Her method was used to perform several trials, and then I used technology to produce many trials. The simulated sampling distribution produced an estimated *p*-value of .011, and all the participants agreed that they should conclude that swimming with dolphins helps relieve depression. As anticipated, the participants did struggle some with designing the simulation. However, for the other steps the participants quickly answered the questions and easily made a correct conclusion based on the simulation.

I have just described how the participants engaged in the simulation tasks as a group. Next, I will discuss the cases of Carrie and Kathleen separately. For each case, I

will share the results concerning their CCK for hypothesis testing both before and after the simulation tasks, which shows how their content knowledge changed after engaging in the tasks. Additionally, I will share how each participant understood simulations and connected simulation and traditional approaches, indicating their SCK.

Carrie

Carrie's story is enlightening and interesting for several reasons. Her background is like many other mathematics teachers in that she did not receive proper training and felt underprepared to teach this subject. Carrie was passionate about being an excellent teacher but knew she lacked the in-depth understanding to truly teach this subject well. She seemed nervous during several sections of the interview when she did not know the answer to some of the questions. She said, "I can't wait to find out all these answers" (Pre-interview, October 17th, 2017). Carrie's story will show how a teacher with several years of teaching experience can still hold many of the same misconceptions that students do, as I will share in the pre-data section. However, her journey will end with enlightenment and show how using simulations for hypothesis testing helped her overcome many of these misconceptions and deepened her understanding of this topic.

I will begin by presenting the data that is used to answer research question number one, "How does engaging in simulation tasks for hypothesis testing influence high school statistics teachers' understanding of traditional hypothesis testing?" To answer this question, it is important to evaluate Carrie's understanding of hypothesis testing before engaging in the simulations tasks and to compare that with her understanding of hypothesis testing after engaging in the tasks. Therefore, in the next

section, I will share what the pre-data revealed concerning Carrie's CCK of hypothesis testing and some of the misconceptions Carrie possessed prior to the simulation tasks.

Research Question One

Pre-data. To obtain an initial assessment of Carrie's CCK of hypothesis testing, I had her complete a pre-open-ended response survey (see Appendix B), the pre-CAOS test (see Appendix C), and a pre-interview (see Appendix D). The theoretical framework for this study identified five KDUs for hypothesis testing. These were the logic of hypothesis testing, probabilistic nature of hypothesis testing, importance of data collection in hypothesis testing, importance of sampling distribution in hypothesis testing, and importance of variability in hypothesis testing. For each of these KDUs, the analytical framework divided them into smaller categories and identified specific understandings to assess. For example, under the logic of hypothesis testing, one category is proof, and the specific understanding to asses is that failing to reject the null hypothesis does not prove the null hypothesis. As described in Chapter Three, I located pieces of data corresponding to the categories and used the understanding listed to assess whether the participant possessed this understanding. By creating analytical memos describing the alignment of the understanding and then synthesizing these notes, I created the descriptive narrative of the pre-data reported in this section. See Table 6 in Chapter Three for a list of these specific understandings for each of the five categories of the theoretical framework. I have organized the pre-data in this section according to these five KDUs. I will begin by addressing the logic of a hypothesis test, which is the first KDU of CCK for hypothesis testing, followed by probabilistic nature, data collection, sampling distribution, and variability.

Logic of hypothesis testing. Carrie was able to explain that indirect reasoning would be used and that two competing hypotheses are needed. She stated, "I introduce it as okay we're going to test a claim that this is actually the proportion or the mean or whatever versus well what if it's not so" (Pre-interview, October 17th, 2017). She also correctly identified the null hypothesis from CAOS test question one. However, when I asked her how to determine if the test was one or two sided, she said, "I struggle with that so much because there's been cases where I thought so it's definitely less than and it's not equal to. Because if the sample is less than the proportion for the null hypothesis, then I'm thinking we should be testing less than, but sometimes it's not that way" (Pre-interview, October 17th, 2017). She did not recognize that the practical needs of the researcher are what determines if the test is one- or two-sided but thought that the sample data should determine this. However, she did seem to understand that a hypothesis test is used to answer a research question about a population from a sample. In the pre-open-ended survey, she was the only participant to acknowledge that a hypothesis test could be used to answer question one. The question asked if a random sample of 20 graduate students indicated that 60% of them were satisfied, would you believe the university's claim that over 80% of all graduates were satisfied. Carrie wrote that to test the claim she "would calculate a test stat, find a *p*-value, and compare to significance level." However, in the same open-ended-survey, when describing a hypothesis test, she wrote that the hypothesis test "will tell us whether the original claim is true or not" (Pre-Open-Ended Response Survey, October 5th, 2017). This indicated that she may not be clear that a hypothesis test does not prove if the null hypothesis is true or not. This same sentiment

was expressed in her pre-interview when I asked her if she had any teaching strategies to help students understand how *p*-values are used to make conclusions. Carrie said,

> Not really, usually they just have a little interactive notebook where they write the formulas and things down, and they have a page in there, like a page for the type 1 and type 2 errors, and a page that talks about when to reject and when to accept, and they just go back and use that as a reference. (Pre-interview, October 17th, 2017)

Not only did this indicate that she would accept the null hypothesis, but this also showed her focus on steps and procedures versus a deeper understanding.

Carrie also indicated that statistical significance does indicate practical significance by her incorrect response to CAOS Tests of Significance question #7, which involved testing a claim with a large sample size and a small difference between the null hypothesis and sample data. The question stated that a newspaper claims that the average age of food stamp recipients is 40 years old. You believe that age is lower and decide to test the claim by taking a random sample of 100 people who receive food stamps and calculate their average age. You find an average of 39.2, which you find to be statistically significant. What is an appropriate interpretation? The correct response was that although the result is statistically significant, the difference in age is not of practical importance. However, Carrie selected that the statistically significant result indicates that the majority of people who receive food stamps is younger than 40.

Finally, the pre-interview revealed that Carrie used the logic that the null hypothesis was what they believed was true instead of focusing on trying to have evidence for the alternative. I used a scenario like in the study by Thompson et al. (2007),

which was described in Chapter Two. For this scenario, I asked Carrie if she believed the claim that people prefer Pepsi over Coca Cola based on evidence from a sample in which 60% preferred Pepsi. I also provided her with a simulated sampling distribution of sample proportions under the null hypothesis that the population was split 50-50 in their soda preference. In the simulation, a proportion of 60% or higher occurred only 2.5% of the time. Carrie said,

> I mean that's just one single sample out of all the possibilities. So, there, of course there's going to be in a sampling distribution, there are going to be some samples that fall outside of the norm. So, I mean that's what this one individual sample is and in that sample there was more people that favorite Pepsi, but no, I wouldn't say in the population. (Pre-interview, October 17th, 2017).

The correct interpretation should have been that the data provides evidence that there is a preference, because obtaining 60 percent or higher would only happen 2.5% of the time if there were not a preference, which is unlikely. However, Carrie indicted that even though it was rare, it could happen. Therefore, it was not evidence that more people from the sampled population prefer Pepsi.

This data showed that Carrie did not understand many of the important components for the KDU of the logic of hypothesis testing. Next, I will share Carrie's understanding for the second KDU of CCK for hypothesis testing, probabilistic nature.

Probabilistic nature. The first component of the probabilistic nature KDU is understanding that a cut-point is necessary when determining whether to reject or fail to reject the null hypothesis and that the probability associated with the sample is used in the decision. She possessed a strong understanding that probability is involved in making

a decision as indicated in her pre-interview when she was describing how to introduce hypothesis testing to her students. She said, "get them thinking that we were finding a probability . . . and we're going to use that probability to determine could it happen again or was it just like a chance occurrence. I just keep it about probability and make sure they understand they're finding the probability" (Pre-interview, October 17th, 2017). However, she seemed to have misconceptions regarding the role of probability. First, she did not compare her *p*-value to an alpha level when working a hypothesis test problem on the pre-open-ended response survey. Additionally, she missed the CAOS test question regarding that the alpha level determines the probability of a type one error. The question asked how many statistically significant results out of 100 would be expected if the null hypothesis were true. Carrie selected zero. Also, in the pre-interview, she was unable to explain why a significance level is set and how to determine the probability of a Type I or II error. The following excerpt from the pre-interview (October 17th, 2017) showed Carrie struggling with these concepts.

Researcher: What's the point of setting a significance level or an alpha level?

Carrie: Since we're coming off of confidence intervals it's something comparable to that. We're not ever going to be a hundred percent certain of anything. So, you want … I don't know how to explain that. I mean if it's a 95% significance level . . . I mean it's .05, that you're comparing the *p*-value to it to, and I mean bigger the *p*-value. Oh no I'm going to confuse myself. I mean the bigger the significance level the more you have for the *p*-value. I mean . . .

I'm confused now. I mean if the significance level was really, really small, then it would be less likely that you would reject the null.

Researcher: So, are there any pros or cons to making the alpha-level smaller or bigger?

Carrie: I don't know if I can explain that off the top of my head. I mean no.

Researcher: Can you determine the probability of making a type 1 or 2 error? Is that possible?

Carrie: I don't know how to do that. So, I don't know.

Although she was correct that a smaller alpha-level would reduce the probability of rejecting the null hypothesis, she did not seem certain, and data collected from the CAOS test and other portions of the interview confirmed her confusion. She was able to select the correct definition of *p*-value on question two of the CAOS Tests of Significance, as the probability of observing an outcome as extreme or more extreme than the one observed if the null hypothesis is true, but then missed how to correctly interpret a *p*-value and when to reject on the other test items about *p*-value. For example, she selected that a large *p*-value is preferred if you wanted to achieve statistically significant results. In the pre-interview, I tried to gain more insight into her understanding as shown in the following excerpt.

Researcher: So, do we reject when we have a small *p*-value or a large *p*-value and why?

Carrie: When it's a small *p*-value, then the probability of that occurring again is very rare, but a large *p*-value for the probability, it's highly possible that it could happen again, so the claim is probably false, so we would reject the null when the *p*-value is large" (Pre-interview, October 17th, 2017).

When looking at all these data pieces, Carrie only possessed the CCK of the correct definition of *p*-value from the components of probabilistic nature of hypothesis testing as indicated in the analytical framework. The KDU of data collection will be shared next.

Data collection. For this KDU, Carrie was unable to recognize the presence of bias in the sampling technique in question 8 of the CAOS Tests of Significance. For this question she indicated that the sample was large enough to provide an accurate estimate of the public's opinion on an issue. However, the data was obtained from Americans who simply mailed in letters stating their opinion, not from a random sample. She knew that a large sample size was preferred but did not recognize that because the sample was not random that a confidence interval should not be constructed. Additionally, although she explained in checking the conditions for a hypothesis test that samples should be unbiased, random, and large enough, she could not explain why. For example, the following excerpt shows our exchange regarding randomness. (Pre-interview, October 17th, 2017).

Researcher: Does it matter if our data comes from a random sample or randomized experiment?

Carrie: Do you mean like the data, like the one little sample. I mean I would assume it needs to come from a randomly-selected or people

> need to be randomly assigned to treatments. I mean all the elements of experimental design need to be in place, but then again, you're just basing your test off that one sample. So, I mean if you're, when you're doing the test, you're seeing if that sample is likely to happen again. So, it may not be a huge deal. So, I just teach that anytime you're doing a sample, it needs to be a random sample.

Carrie did not seem to understand the difference between a random sample and a randomized experiment or why it is important. The final two KDUs of CCK will be shared next.

Sampling distribution and variability. The last two KDUs of the importance of sampling distribution and variability will be reported together in this section. Carrie correctly answered all the questions concerning sampling distribution and variability on the CAOS Sampling Variability test, except for question three, which incorporated sample size and probability. She selected that two hospitals, one with an average of 10 births per day and one with an average of 50 births per day, were equally likely to record 80% or more female births. The problem stated to assume that half of all newborns are girls. She did not recognize that the larger sample size would result in a sampling distribution in which obtaining a proportion of 80% or higher would be less likely due to decreased variability. However, on the pre-open-ended survey and the interview, she was able to provide correct definitions for sampling distribution and did acknowledge that samples would vary. Also, although not clearly stated, she seemed to know that there is a difference between the three levels of population, sample, and sampling distribution.

However, she struggled with describing the specific characteristics of a sampling distribution and its corresponding variability as shown in the following excerpt (October 17th, 2017).

Researcher: Can you explain the difference between population, sample, sampling distribution and also think about the variability of each or is there a difference?

Carrie: Okay, so obviously the sample is a group that is chosen out of the population to represent the whole population. The sampling distribution is every possible sample of a certain size that could be taken out of the population and as sample size increases you're going to have less variability in your distribution. Okay, so the sample is going have a lot of variability. Sample compared to . . . So, a sample is going to have a lot of variability. The bigger the sample size the smaller the variability, but a sample is going to have a lot of variability than a sampling distribution is going to have, since it's all possible samples, it's going to have less variability than one sample and as you take bigger and bigger samples the variability will decrease.

She stated that the sampling distribution is constructed by looking at all possible samples, but she did not articulate that the sampling distribution is the distribution of all possible statistics, which have been calculated from the samples. She was also a little unsure in her description of variability. She correctly stated that as you take bigger and bigger samples that the variability will decrease, but she did not explicitly state that this

was the variability of the sampling distribution. She may have simply memorized that larger sample results in decreased variability.

In conclusion, the pre-data indicated that Carrie possessed a rudimentary procedural understanding of hypothesis testing. She could only completely work out a one sample hypothesis test problem if she had access to a textbook, and even with a textbook, she could not complete the two-sample hypothesis test problem given on the pre-open-ended response survey. In the pre-interview, she stated that she did not cover two sample hypotheses testing in her classroom. Regarding terminology and procedures, she could provide correct definitions and explain some of the steps that should occur for hypothesis testing. However, when attempting to answer questions or provide explanations which required a more in-depth understanding, she was unable to do so. For example, although she knew that samples should be unbiased, she did not identify when bias was present. Also, when she did not have the steps in front of her, she incorrectly stated that you should reject the null when the *p*-value is large. She believed a large *p*-value meant that it is likely that your data could happen again, so you would reject the null hypothesis. Similarly, she could define sampling distribution and understood that larger samples are better. However, she could not apply this concept to recognize that smaller samples are more likely to be farther from the mean of the sampling distribution, which would be equal to the true value of the population parameter. Now, to see what changes occurred in her CCK for hypothesis testing, I will share the post data.

Post-data. After completing all three simulation tasks, Carrie completed the post-COAS, post-open-ended question survey, and post-interview. I used the analytical framework for this study to compare the data with the list of understandings for each

KDU and assess whether Carrie possessed this understanding. I also compared this with her pre-data to determine changes in Carrie's understanding of hypothesis testing after engaging in the simulation tasks. I then synthesized the notes of these comparisons to create the descriptive narrative of Carrie's post-data, which I organized based on the five KDUs of CCK from the analytical framework, as was done in the pre-data section.

Logic of hypothesis testing. The pre-data indicated that Carrie understood that two competing hypotheses are needed for a hypothesis test. However, this understanding seemed to be strengthened by the simulation tasks. In the post interview, I asked her to use a simulation to test the claim by a university that at least 80% of all graduate students are satisfied. The result from a sample showed that only 60% were satisfied. To begin solving the problem, Carrie wrote $p \geq .8$ for the null hypothesis and $p < .8$ for the alternative. She correctly typed in $p = .8$ for the null hypothesis when setting up technology to run the simulation. Later in the interview, when we discussed the null and alternative from a textbook problem, Carrie noticed the null hypothesis used equal to and realized that she did not write out her null hypothesis for the simulation as equal to. She asked, "Why did I not put equal to for this problem". I explained that the worst-case scenario was equal to .8 and that we needed a specific sampling distribution to construct. Carrie realized, *"*Because I mean it doesn't affect how you set it up. We did do the calculations based on $p = .8$. So even if it is even more than .8 then it's going to be even more surprising." This showed she had an even greater appreciation for the role of the null hypothesis and the idea of creating competing hypotheses versus simply following the procedures in a textbook.

However, Carrie still did not understand how to determine if the test should be one or two-sided. Each simulation task was based on a one-sided test, so this was not surprising. Additionally, Carrie still missed the question on the CAOS test about practical significance. This was also not surprising because the lesson plan did not focus on determining if statistically significant results were also practically significant.

The simulations did seem to help Carrie realize that she could have made a mistake, such as a Type I or II Error. On the post-CAOS test, for the question which asked how many statistically significant results would be expected out of 100, with an alpha-level of .05, if the null hypothesis was true, she correctly switched her answer from zero to five. When I asked her, what made her change her mind, she said that before the simulations that she had not really thought about making a mistake. She said, "I really haven't been considering the whole Type 1 and Type 2 error in the back of my mind, that I haven't been thinking about, oh this could still be wrong" (Post interview, December 18th, 2017). Therefore, Carrie had developed a better of understanding of the CCK of the logic of hypothesis testing by acknowledging that a mistake could have been made. The tasks also helped Carrie develop the KDU of probabilistic nature, which will be discussed next.

Probabilistic nature. For this KDU of hypothesis testing, the pre-data indicated that Carrie could only correctly define *p*-value and did not possess the other CCK of this category listed in the analytical framework. However, after completing all three simulation tasks, the post-data indicated that this category was the most influenced by the tasks. Carrie even acknowledged that she felt that simulations helped the most with understanding this component when she wrote, "Students are going to better understand

the *p*-value and the reject/fail to reject a lot better" (Post-open-ended question survey, November 27th, 2017). Both the pre-CAOS and pre-interview showed that Carrie believed that you rejected the null hypothesis if the *p*-value was large. She marked the correct response on the post-CAOS and explained the concept of when to reject correctly in the post-interview. She explained, "You reject the null if the *P*-value, which is probability, is less than the significant level. If it was just very low, then it just didn't happen by chance. This was a legitimate thing." (Post-interview, December 18th, 2017). Although she said it did not happen by chance instead of it was unlikely to happen by chance if the null was true, she indicated that she was starting to understand why a small *p*-value results in rejecting the null hypothesis. To probe her understanding more, I asked her why she rejected the null hypothesis in the simulation she just conducted, concerning the graduate approval rating. She said, "Because if we did our sampling procedures correctly that just probably wouldn't have happened." (Post-interview, December 18th, 2017). This time, she did not say that it would not have happened if the null were true but indicated that it probably would not have happened if the null were true, which gave her evidence to conclude the alternative.

Also, Carrie showed signs that she was beginning to understand the point of setting an alpha-level and that this corresponded to the probability of a Type I error. In the post interview, when conducting her own simulation and concluding that her *p*-value meant to reject the null hypothesis, she said, "I mean I'm using five percent as guidelines, since we weren't really told" (Post-interview, December 18th, 2018). This contrasted with her pre-data when she never mentioned comparing the *p*-value to anything. Also, when I asked her the probability that she incorrectly rejected the null hypothesis, she correctly

said five percent. Her understanding of alpha being the probability of a Type I error was confirmed on the post-CAOS test, where she indicated that she would expect about five percent of tests to be statistically significant if the null hypothesis were true and the alpha level was .05. The previous two KDUs of the logic of hypothesis testing and probabilistic nature were the most influenced by the simulation tasks. I will share the final KDUs of hypothesis testing next.

Data collection. For this KDU, Carrie only missed the pre-test CAOS question concerning the identification of bias to determine that results of a hypothesis test were invalid. On the post-test she answered this question correctly. She did not remember why she changed her answer when I asked her about this question in the post-interview. Therefore, it is unclear if she simply made a careless mistake or if something about the simulations allowed her to develop this understanding. The pre- and post-data indicated that she possessed the CCK for the remaining aspects of this category.

Sampling distribution and variability. Carrie had answered most of the questions on the pre-CAOS and pre-opened-ended question survey correctly for these KDUs. Her pre-interview also indicated that she understood most of these concepts. However, she correctly answered one of the post-CAOS test items for this KDU that she had previously missed, regarding her understanding about larger samples resulting in a sampling distribution with less spread. The question asked which hospital, one with an average of 50 births per day and one with an average of 10 births per day, was less likely to record 80% or more female births. Also, the problem said to assume that 50 percent of newborns are girls. On the pre-test, she incorrectly selected that the probability was the same, instead of selecting that the hospital with an average of 10 births per day is more likely,

because the resulting sampling distribution would have more spread. However, she answered this question correctly on the post-test, but she could not tell me why she had changed her answer in the post-interview.

Although Carrie did not know why she changed her answer on the post-test for the category of sampling distribution and variability, she did comment that "the simulation helped me see how the sampling distribution really does start to look normal" *(*Post-open-ended question survey, November 27th, 2017). So, even though the pre-data showed that she could correctly define terms and answer questions regarding sampling distribution and variability, the tasks helped her visualize the sampling distribution better. This may have been the reason she was able to correct the question regarding sample size and the sampling distribution that she had previously missed on the pre-CAOS test for this KDU.

In conclusion, the post-data indicated that Carrie had developed her CCK for hypothesis testing and understood how to use simulations to conduct a hypothesis test. For the CAOS assessment, she answered all the questions correctly that she had missed the first time, except the question regarding practical significance. The simulations did not help her understand that even though statistical significance is achieved, that does not imply that the results are meaningful in the real world. The questions she missed the first time and correctly answered on the post-test concerned *p*-values, types of errors, and sampling distribution. She identified the correct interpretation of *p*-value and knew when to reject. Additionally, she correctly identified Type I and II errors and knew that the probability of a Type I error was the alpha-level.

The post-interview also indicated that Carrie had developed her CCK for hypothesis testing through her engagement in simulation tasks. She was able to correctly explain that a small *p*-value was when you reject the null hypothesis. Additionally, she could correctly interpret results from a provided simulation and was able to conduct her own simulation to answer a question that would typically be answered by a hypothesis test.

When asked how these tasks influenced how Carrie thought about hypothesis testing, she wrote, "Everything! The whole process became clear to me from writing the hypothesis to determining what the conclusion would be" (Post-open-ended question survey, November 27th, 2017). Additionally, in the post-interview, when talking about the knowledge she gained from the simulations, she said, "I mean obviously this helped my conceptual understanding and my ability to explain to someone else. I think before I was like OK I know what steps we have to do, and I'm really good at it, like doing those steps, but now I actually understand what we're doing and why we're doing it and what it means at the end" (Post-interview, December 18th, 2018).

The previous sections focused on identifying specific KDUs of Carrie's CCK for hypothesis testing that Carrie possessed prior to engaging in simulations tasks and then which ones were influenced by the simulation tasks, by analyzing the post-data. This served to answer research question one by showing how her CCK was influenced by the tasks. Next, I will finish answering research question one by sharing the overall themes for this question, which were the same for Kathleen.

Themes. A deductive analysis, as described in Chapter Three, was used to determine the changes in Carrie's CCK that were just described. An inductive analysis

was used to determine the influencing factors for these changes. From this analysis, two overall major themes emerged regarding research question one. My first theme was that changes in Carrie's understanding of hypothesis testing were influenced by the focus of the simulation tasks on concepts and logic instead of procedures, which develops the KDU of the logic of hypothesis testing. My second theme was that simulation approaches focus on visualizations, which help develop the KDU of probabilistic nature. Each of these themes will be described next.

Research Q1, Theme One. Simulation tasks focus on concepts and logic instead of procedures, which develops the KDU of the logic of hypothesis testing. From the pre-data when Carrie described the logic of a hypothesis test, she wrote,

> We will use the information available to calculate a test statistic, which is similar to a *z*-score. We will use that *z*-score to identify the *p*-value, which tells us the probability of this event occurring again, assuming the original claim is true. (Pre-Open-Ended Response Survey, October 5th, 2017)

This showed that Carrie focused on procedures when discussing a hypothesis test. This focus was also seen in the pre-interview, as shown in the following excerpt.

> Researcher: Why do you think that this subject or topic is so hard for students? What do you think about it that makes it so difficult?
>
> Carrie: I think it's because there are so many steps to it, and they don't understand what they're doing on each of the steps. They don't have the conceptual understanding of it, and a lot of times I think the teacher doesn't really understand it either, because we weren't taught it. So, we're just teaching the steps to go through, but then

there's a lot of what ifs and things that could derail some of the steps. And it just becomes a big chaotic mess sometimes, and they don't know quite what to do.

Researcher: So, is there like certain prerequisite knowledge from the other chapters that you think are critical that they understand before they can grasp the concept?

Carrie: I mean yeah. You have to understand the normal distribution. You're going to have to understand a *z*-score. Have to understand the idea of statistically significant. Of course, meeting the conditions . . . the normal condition and the large counts . . . the 10% rule. All of those. Knowing how to calculate mean and standard deviation.

Carrie mentioned several times about steps and calculations, but she did not discuss the logic behind a hypothesis test. However, during the tasks, we never calculated a *t*-test statistic and obtained an approximate *p*-value from the simulated sampling distribution. The tasks were specifically designed to focus on the logic behind a hypothesis test, not the procedures.

In one of the analytical memos from the pre-data and another from Task A, I had commented that Carrie seemed to focus on procedures. However, I had not made this notation for the other tasks and the post-data. Additionally, as I was constructing the task narratives, I noticed that Carrie was focusing less on procedures. When comparing the data from each of the tasks, I noticed a trend in question number two. Question two had the participants list all the possible hypotheses for the given scenario. Carrie's responses

showed her moving away from using traditional hypothesis test notation, to focusing on the hypotheses in context, as shown in Table 11 .

Table 11

Comparison of Tasks' Question Two

Task A	2. How many different possible hypotheses could we make for this situation regarding pre-verbal children and their choice of a toy? What are they? $p=.5$ $p>.5$
Task B	2. How many different hypotheses could we make for this situation regarding the averages of scores of students who take the exam on yellow paper and on white paper? What are they? White > Yellow $w=y$
Task C	2. What are all the different hypotheses that we could make for this study? Swimming with dolphins relieves depression. Number of improvers equal.

For task A, she wrote $p = .5$ and $p > .5$, for the null and alternative, relying solely on traditional hypothesis test notation. For task B, she mixed the context with traditional notation by writing Yellow = White and Yellow > White for the hypotheses. By task C, she had completely moved away from the formal notation and wrote out "Swimming with dolphins relieves depression" and "Number of improvers equal" for the two hypotheses. She was no longer relying on the traditional notation, but instead focused on what we were trying to logically deduce from the data.

Carrie's transition to focusing on concepts and logic versus procedures was also seen in her response in the post-data, which asked her to describe the logic of a hypothesis test again. In the pre-data she focused on calculations, as noted in the

beginning of this section, but on the post-open-ended question survey, she wrote, "We want to see whether our trial/experiment was a fluke; what are the chances of that happening again?" (November 27th, 2017). She did not mention steps and procedures at all. This statement is particularly important because it shows that Carrie understood Big Idea Three from *Developing Essential Understanding of Statistics* (Peck et al., 2013), regarding the logic of hypothesis testing. As mentioned in Chapter Two, this idea is that a hypothesis test is used to answer the question, "Do I think that this could have happened by chance."

Her post-interview also showed how her KDU of the logic of hypothesis testing was strengthened by the focus of simulation tasks on concepts and logic. When I asked her to complete a problem using a simulation approach, she wrote out $p \geq .8$ for the null hypothesis and $p < .8$ for the alternative. With a traditional test, the alternative would be equal to, but she was focusing more on the logic of setting up competing hypotheses versus following the traditional steps of using "=" for the null hypothesis. All the tasks had the participants think through the hypotheses logically instead of trying to simply state a null hypothesis in terms of equal to and then write out the corresponding alternative. Carrie was thinking through the problem instead of relying on the steps. Next, I will share the second research theme.

Research Q1, Theme Two. The second theme was that simulations focus on visualizations, which help develop the KDU of probabilistic nature. Carrie's pre-data indicated that this KDU was where Carrie lacked the most knowledge. She had missed almost every related question on the pre-CAOS and pre-open-ended response survey and incorrectly explained concepts for this KDU. However, this was also the KDU that saw

the most change after the simulation tasks, as reported in the post-data section. This theme was not originally part of my results for Carrie. However, after completing Kathleen's case, I revisited Carrie's data as a way of double-checking my results. This theme was apparent for Kathleen, as will be shown in her section; however, Carrie did not explicitly mention the visualization component as often as Kathleen did. However, through rereading Carrie's data, I believe that the visualization component is what helped Carrie develop her understanding of this KDU as well. I found four main pieces of evidence to support this conclusion, as shown in Table 12.

Table 12

Carrie's Evidence for RQ1, Theme Two

Data Piece	Evidence
Post-Task B Reflection	" a 'real' normal curve rather than the one in our books" (November 15th, 2017)
Post-Task C Reflection	"Use technology to simulate doing this many, many times *see Normal distribution" (November 20th, 2017)
Post-Open-Ended Question Survey	"The simulation helped me see how the sample distribution. Really does start to look normal!" (November 27th, 2017).
Post-interview	"So, the simulation just gives us a quick and easy way to see the visualization of the distribution and to quickly find the probability of that value or less happening" (December 18th, 2017).

The first piece of evidence for this theme was found in the post-task B reflection when she was asked to connect the traditional and simulation approaches. This shows how the simulation approach allowed Carrie to focus on seeing a simulated sampling distribution versus the theoretical one described in textbooks. However, it is interesting to note that Carrie referred to the simulated sampling distribution as the 'real' normal curve, instead of the theoretical one in the textbooks as the real curve. I did not notice this when I first read the data and did not follow-up in the post-interview. I can only infer what made her make this comment. It could be that she thought the simulated sampling distribution was more real because it was produced by using the actual data, and she was able to see the construction of the distribution on the screen. Also, one could argue that Carrie had never made sense of the theoretical distribution in the textbook and that it why it did not seem real to her. This is an important element, because the main difference in the two approaches is the type of model used for the sampling distribution. For the simulation approach, the model is created through simulation, but for the traditional approach, the model is the theoretical sampling distribution.

Additional evidence of the visualization aspect was found in the post-task C reflection when Carrie emphasized visualization by starring where she wrote "see normal distribution". The ability to see the distribution constructed herself versus what she may have viewed as an "unreal" normal curve in the textbooks was powerful for her. Finally, she mentioned being able to "see" the sampling distribution in the post-open-ended survey, and her quote in the post-interview gives more proof that the visualization component is what helped her understand the *p*-value better, as she explicitly mentions

this aspect. Her quote expressed the idea that simulations are easier to understand and that she could see the p-value instead of having to calculate it.

In conclusion, the data in this section answered research question one by sharing what aspects of Carrie's CCK changed after engaging in the simulation tasks. Additionally, the inductive analysis of the data was shared, which determined the themes for research question one. In the next section, the data which answered research question two will be shared, along with the theme for research question two.

Research Question Two

Research question two was, "How do simulation tasks influence high school statistics teachers' understanding of simulations and how do they make connections between traditional and simulation approaches for hypothesis testing?" Most of the data pieces that answered this question corresponded to the post-task reflections; therefore, a descriptive narrative of the post-task reflections will be shared in this section. This data was analyzed deductively using the analytical framework described in Chapter Three, and I will share my tables used to compare Carrie's simulation models and connecting approaches in this section as well. Additionally, an inductive analysis was used to determine the theme corresponding to research question two, which will be presented at the end of this section.

After task A, I asked Carrie to draw a diagram representing the simulation task, work a traditional hypothesis problem correlating to the simulation task, and to connect the simulation and traditional approaches. The following diagram (see Figure 9) shows Carrie's model of the simulation task.

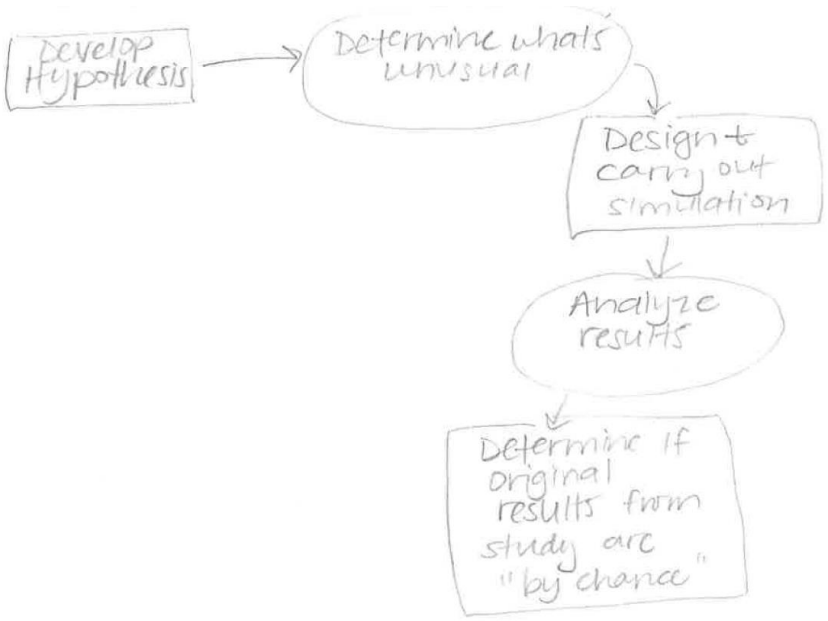


Figure 9. Carrie's Simulation Model

Carrie's model shows a five-step approach to conduct a hypothesis test through simulation. Carrie referenced her task worksheet while drawing her diagram. I compared her steps to the modified Lane-Getaz and Zieffler's (2006) SPM, which was listed in my analytical framework in Chapter Three to assess the participant's understanding of simulations by comparing it to a model representative of the literature. In Table 13, I listed the modified steps of the SPM on the left and aligned them to Carrie's steps on the right.

Table 13

Alignment of Carrie's Simulation Model Post-Task A with Lane-Getaz and Zieffler's (2006) modified SPM

Lane-Getaz and Zieffler (2006) SPM	Carrie's Model
1. Establish population parameters	1. Develop hypothesis 2. Determine what is unusual
2. Generate samples through simulation	3. Design and carry out simulation
3. Create sampling distribution and assess unusualness.	4. Analyze Results 5. Determine if study's results are "by chance"

Step one is similar but described using different terms. Carrie named this step develop hypothesis instead of establishing population parameters. However, when you write a hypothesis, you are writing it in terms of your population parameters. Therefore, step one can be considered the same. Carrie's step two of determining what is unusual before designing a simulation was not part of any of the simulation models discussed in Chapter Two. However, it was part of the lesson plan design to have participants think about anticipated variability before revealing the study results and to consider how much evidence would convince you to reject the null hypothesis, like setting an alpha-level in a traditional test. Therefore, her inclusion of this step was most likely due to the lesson plan design. Step two from the SPM corresponded to Carrie's third step. Again, just slightly

different wording was used. For step three, Carrie did not mention creating a sampling distribution like the SPM does. However, her last two steps of analyze results and determine if the study's result are by chance is the same thing as the SPM's second part of step three of assess unusualness.

Next, Carrie completed a one-sample traditional hypothesis problem, which corresponded to the task (see Figure 10). I had her work this problem so that she had a traditional hypothesis test problem to compare to the simulation approach. The instructions asked her to complete each step and explain why each step was important. She wrote out $p = .5$ but did not label this as the null hypothesis. She also did not identify the alternative hypothesis. She correctly calculated the *z*-test statistic and *p*-value. She explained that the *z*-score allowed one to find the probability of this event occurring and used a *z*-table to determine the probability. She explained that the *p*-value gave the probability of this event happening by chance but did not explicitly state that this probability was based on the null hypothesis being true. She did make a correct conclusion.

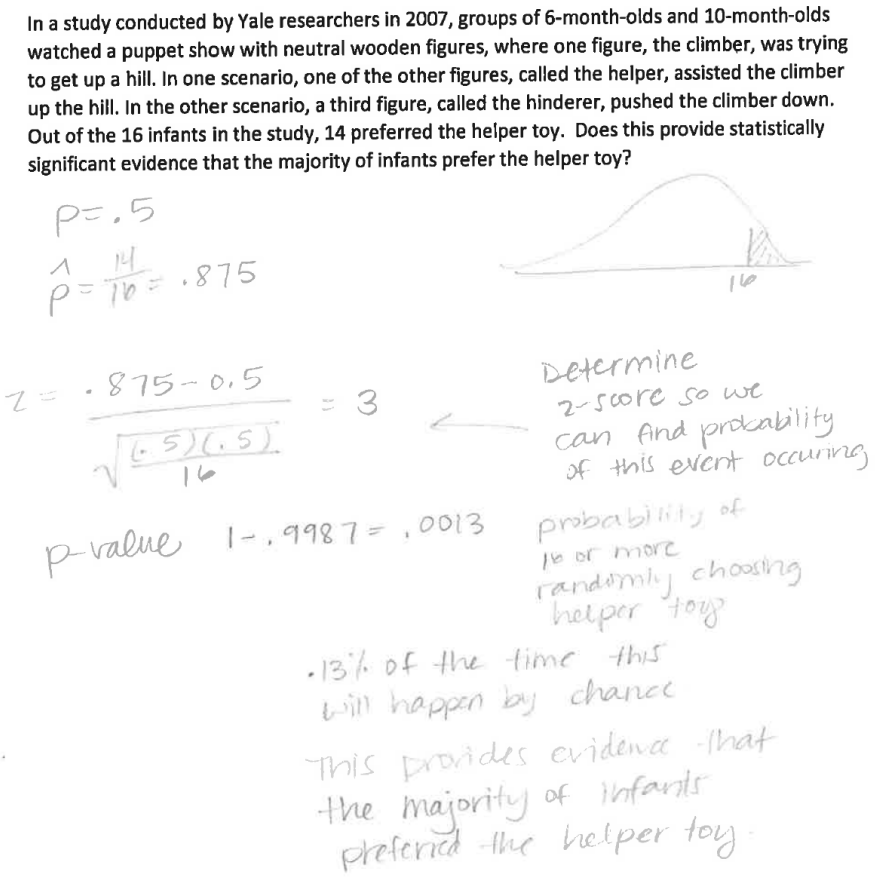


Figure 10. Carrie's Post-Task A Reflection Question Two (November 9th, 2017)

The third piece of post-task A reflection was intended to see how Carrie made connections between her simulation model that she created and the traditional hypothesis test problem that she worked. The question asked, "Using your diagram, explain how you see the simulation task connected to the traditional hypothesis test problem that you worked in the previous section." She wrote,

> When we calculated the *p*-value, we were finding the chances of that event happening again, which was done much easier through the simulation. The

> simulation seems to take some unnecessary steps out of the process. (November 9th, 2017)

She did not explicitly make connections between the steps of the two approaches as I had intended. Instead, she provided a general comparison. She saw the simulations as easier and less complicated to arrive at a conclusion.

The next post-task reflection was after task B, concerning the yellow-white exams. In this reflection, instead of having Carrie draw a diagram, I asked her to create a step by step guide that students could use to conduct a hypothesis test using simulations. I asked her to list steps instead of the diagram to see if I could obtain more details about her thinking about the simulation design. I again allowed her to have the task handout to reference as she completed this. The following shows how Carrie listed the steps.

1. Form various hypothesis about the scenario.
2. Decide what the difference should be if the treatment has no effect.
3. Determine how much of a difference would be surprising if the treatment has no effect.
4. Give the actual difference from experiment.
5. Perform a simulation just to generate some possible "chance" differences.
6. Use technology to simulate taking many samples and calculating the difference.
7. Calculate what % of the time the difference obtained occurs.
8. Determine if this % is high enough to be "surprising" or if it is due to chance.

Notice in step seven that Carrie wrote, “determine if this % is high enough to be “surprising” or if it is due to chance” (Post-Task B Reflection, November 15th, 2018). I did not recognize until after data collection was complete that she wrote high enough instead of small enough to determine if something is surprising. Therefore, I did not ask a follow-up question regarding this statement in the post-interview. Carrie may have still been struggling with whether a large or small *p*-value results in rejecting the null hypothesis. However, this may have just been a mistake, because Carrie correctly interpreted the *p*-value on another section of the post-task B reflection, which is shown later. Like Carrie’s simulation diagram from the post-task A reflection, these steps seemed to be directly linked to the lesson plan design. She also expanded her steps from five to eight. Steps two through five are not necessary for a simulation approach but were included in the lesson plan design for pedagogical reasons. If you omit these steps, then you can see more of an alignment to Lane-Getaz and Zieffler’s (2006) modified SPM as shown in Table 14.

Table 14

Alignment of Carrie's Simulation Model Post-Task B with Lane-Getaz and Zieffler (2006) Modified SPM

Lane-Getaz and Zieffler (2006) SPM	Carrie's Model
1. Establish population parameters	1. Form various hypotheses about scenario.
2. Generate samples through simulation	6. Use technology to simulate taking many samples and calculating difference.
3. Create sampling distribution and assess unusualness.	7. Calculate what % of the time the difference obtained occurs. 8. Determine if this % is high enough to be "surprising" or if it is due to chance.

Additionally, the post-task reflection asked Carrie to interpret the results of a traditional hypothesis test concerning the same scenario, using the terminology of probability, *p*-value, and statistically significant. Carrie wrote,

> A difference of 6.3 points yields a high *p*-value. There is a 13.77 percent chance of obtaining a difference greater than or equal to 6.3, assuming there is no actual difference in performance on yellow paper. This tells me that it could easily be a fluke and that the yellow paper probably has nothing to do with it. (Post Task B Reflection, November 15th, 2017)

Her response showed a correct interpretation of the *p*-value and indicated that a *p*-value is calculated based on the null hypothesis being true. This shows that her understanding of *p*-value is changing. Additionally, she stated that the result could easily be a fluke, indicating a recognition that a *p*-value is a calculation based on

chance occurrence. The word "probably" that she used in her conclusion could be viewed as both a positive and negative. On the positive side, it shows that Carrie understood that her conclusion may be wrong. However, the word probably is also associated with probability, which indicated that Carrie may possess the misconception that a small *p*-value means that the null hypothesis has a small probability of being true versus the correct interpretation as the conditional probability of obtaining results as extreme or more extreme *given* the null hypothesis is true.

The final piece of post-task B reflection asked Carrie to connect the steps of a traditional hypothesis test to the simulation approach. This time I asked her to try to provide more details than the first time and to be sure to explain how they are connected. Although more details were given, Carrie still did not explicitly link all the steps of the two approaches. She wrote,

> We are using a simulation to estimate a *p*-value. As we take more samples, our *p*-value approaches the traditional calculation. We are finding area under the normal curve, which is similar to calculating a *t*-value and finding probability/ area of a higher value using the tables. We just have a "real" normal curve rather than the one in our books. We created the curve through simulation. (November 15th, 2017).

Instead of explicitly listing steps, Carrie provided a general overview and did not comment on steps of the simulation that she previously mentioned, such as determining the hypotheses. The main connection that she made was that the simulated sampling distribution replaced the step of calculating the test statistic. In comparing her reasoning

to Lane-Getaz and Zieffler's (2006), her explanation merges steps three, four, and five and omits the other, as shown in Table 15.

Table 15

Carrie's Comparison of Connection of Traditional and Hypothesis Test Steps

Lane-Getaz and Zieffler (2006)	Carrie's Connections
Step3: Simulation: Select the appropriate summary statistic Hypothesis Test: Calculate *z* or *t* test statistic	Simulation: Compile summary statistic for distribution formed by simulation and find area under the curve.
Step4: Simulation: Compile summary statistic for distribution formed by simulation Hypothesis Test: Graph the theoretical sampling distribution based on a function	Hypothesis Test: Calculate *z* or *t* test statistic and find *p*-value from tables.
Step5: Simulation: Assess rareness by finding observed data on simulated sampling distribution and calculating approximate *p*-value. Hypothesis Test: Assess rareness by large or small *p*-value OR large or small test statistic.	

The final task was the swimming with dolphins' scenario. After the task, to offer Carrie the chance to be more explicit about how the simulation and traditional approaches were connected, I did not have her list the steps of each approach on separate pages like I did for the previous two tasks. Instead, I asked Carrie to list all the steps for a traditional hypothesis on the left-hand side of a paper, list the steps for a simulation approach on the right-hand side, and then to draw arrows connecting the two approaches. I did not allow her to reference her task handout when she did this. I also emphasized that she should

focus on what were the essential steps of each approach and to not focus on what the steps were from the lesson. I hoped this would provide an opportunity for her to focus on the overall big picture of using simulations versus listing steps based on the lesson plan design. Figure 11 shows her response.

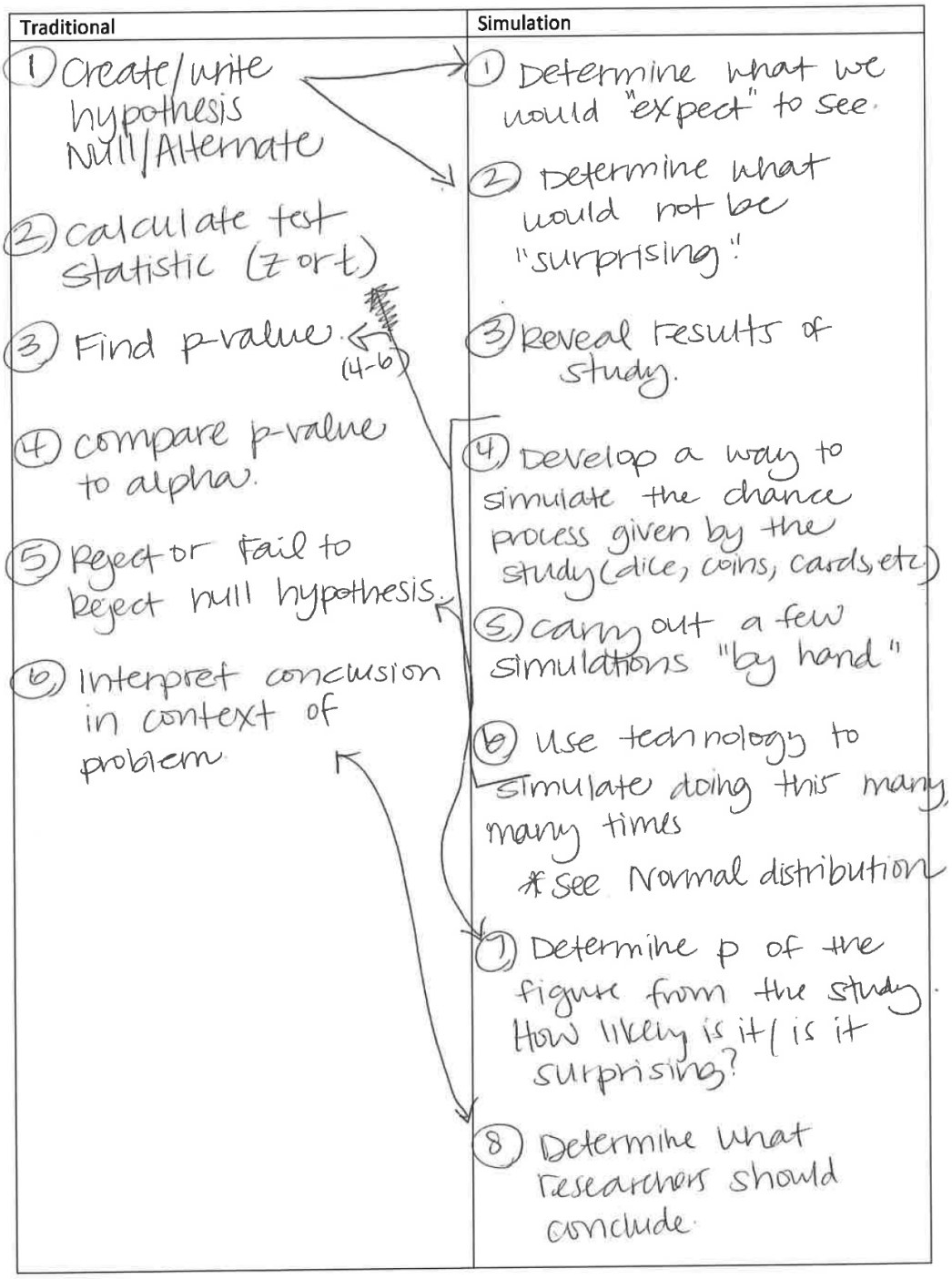


Figure 11. Carrie's Connection of Simulation and Traditional Approach Post Task C

Carrie's left-hand side of her diagram shows how she listed the steps of a simulation approach. By only using the left-hand side of her figure, which corresponds to the simulation approach, I will show how this approach aligns with Lane-Getaz and Zieffler's (2006) modified SPM model first, as shown in Table 16.

Table 16

Alignment of Carrie's Simulation Model Post-Task C with Lane-Getaz and Zieffler (2006) Modified SPM

Lane-Getaz and Zieffler (2006) SPM	Carrie's Model
1. Establish population parameters 2. Generate samples through simulation 3. Create sampling distribution and assess unusualness.	1. Determine what we would "expect" to see. 6. Use technology to simulate doing this many, many times *see Normal distribution 7. Determine *p* of the figure from the study. How likely is it/ is it surprising? 8. Determine what researchers should conclude.

How Carrie listed the steps of a simulation approach was consistent in the post-task A and B reflections, with slightly different wording used. Additionally, the same alignment between Carrie's model and the SPM was found when constructing the tables. I did not list Carrie's steps two through five, because they are not represented in Lane-Getaz and Zieffler's (2006) SPM. These are the steps that were aligned with the lesson plan. Next, I will compare how Carrie aligned the simulation and traditional steps with Lane-Getaz and Zieffler's (2006) connections (see Table 17). I have omitted the

numbering to avoid confusion. Also note, that Lane-Getaz and Zieffler's (2006) model for connecting approaches lists five steps for the simulation when connecting approaches in comparison to their three tiers of the SPM.

Table 17

Alignment of Lane-Getaz and Zieffler's (2006) Connecting Approaches with Carrie's

Traditional Approach		**Simulation Approach**	
Lane-Getaz and Zieffler (2006)	Carrie	Lane-Getaz and Zieffler (2006)	Carrie
Statement of null and alternative hypothesis	Create/write hypothesis. Null/alternative	What if scenario? Determine a model.	Determine what we would "expect" to see. Determine what would not be surprising
Check conditions	Not listed.	Repeat the random sampling or random assignment through simulation.	Not listed.
Calculate *z* or *t* test statistic	Calculate test statistic (*z* or *t*)	Select the appropriate summary statistic.	Not connected to a simulation step.
Graph a theoretical sampling distribution based on a function	Not listed.	Compile summary statistic for distribution formed by simulation.	Not listed.
Assess rareness by large or small *p*-value OR large of small test statistic	Find a *p*-value.	Assess rareness by finding observed data on simulated sampling distribution and calculating approximate *p*-value.	Develop a way to simulate the chance process given by the study (dice, coins, cards, etc.)
			Carry out a few simulations "by hand."
			Use technology to simulate doing this man, many times. * see Normal distribution
	Compare *p*-value to alpha		Not connected to a simulation step.
	Reject or fail to reject the null hypothesis.		Determine *p* of the figure from the study. How likely is it/is it surprising?
	Interpret conclusion in context of problem.		Determine what researchers should conclude.

Both step one connections involving the null and alternative hypothesis are similar. However, Lane-Getaz and Zieffler (2006) included checking conditions and graphing the theoretical sampling distribution, but Carrie did not. Additionally, Carrie listed the step of calculating a *z* or *t* test statistic but did not connect this traditional step to any steps in the simulation approach. Finally, Carrie gave much more emphasis to the *p*-value and making a conclusion. Lane-Getaz and Zieffler (2006) only included one step for this in both the traditional and simulation approaches, but Carrie divided this last step into four for the traditional approach and five for the simulation approach. Carrie did not possess a strong understanding of *p*-value prior to the simulation tasks, and this piece of CCK regarding *p*-value was highly influenced by the tasks as reported in the post-data section. Therefore, Carrie may have placed more emphasis on this because her understanding of this step was the most influenced. Next, I will share the theme for research question two.

Theme. In terms of research question two, the previous sections focused on how Carrie understood simulations and how she connected the simulation approach to the traditional approach. For this research question, one major theme emerged. The theme was that the lesson plan design influenced how Carrie understood simulations and connected approaches.

Research Q2, Theme One. Understanding of simulations and the connection of simulation and traditional approaches was influenced by the lesson plan design. After task A, Carrie's flowchart showed the steps of a simulation as 1) develop hypothesis, 2) determine what's unusual, 3) design and carry out a simulation, 4) analyze results, and 5) determine if original results from study are "by chance". By the last task, Carrie had

expanded the simulation steps to eight. These eight steps could be directly linked to the steps of the lesson plan as seen in Table 18. The left hand-side of the table lists Carrie's simulation steps reported after task C. The right-hand side shows the questions listed in the task C handout. Although the questions are slightly different for each task, based on the context, the purpose of each question was the same for each task. I kept the numbering consistent for steps used by Carrie and the lesson plan; therefore, numbers do not align across the table.

Table 18

Carrie's Steps connected to Lesson Plan

Carrie's Steps	Lesson Plan
1. Form various hypotheses about the scenario.	2. What are all the different hypotheses that we could make for this study?
2. Decide what the difference should be if the treatment has no effect.	3. a. In statistics, we typically subtract two numbers in order to compare them. If the null hypothesis is true, what would be the most likely outcome (difference in number of improvers between the dolphin group and the control group)?
3. Determine how much of a difference would be surprising if the treatment has no effect.	3. b. Still assuming that the null hypothesis is true, what kind of results (difference in the number of improvers) would you not be surprised to see when this study is conducted with 30 participants?
4. Give the actual difference from the experiment.	4. In the actual study, Antonioli and Reveley found that _____ of 15 subjects in the dolphin therapy group showed substantial improvement, compared to _____ of 15 subjects in the non-dolphin (or the control) group. Complete the table based on these results.
5. Perform a simulation just to generate some possible "chance" differences.	5. Determine a plan for simulating the study.
6. Use technology to simulate taking many samples and calculating the difference.	6. b. Now, we will use technology to simulate this experiment many, many times under the assumption that the null hypothesis is true. Based on this simulation how surprising are the actual results of this study? Explain your reasoning
7. Calculate what percent of the time the difference obtained occurs.	6. c. Based on the results of the simulation, how likely is a difference of __ or greater? Explain.
8. Determine if this percent is high enough to be "surprising" or if it is due to chance.	6. d. Based on our simulations, what conclusion should the researcher draw? Justify your conclusion.

The only part of the lesson plan that is not explicitly mentioned is question one where the participants give their own guess about the results of the experiment, and question 6a, which asked participants to comment on the results of the initial simulation without technology. Her explanation of a simulation approach went beyond the traditional three steps found in the literature and can be directly corresponded to the lesson plan design as seen in the table. As noted earlier, Carrie did write high enough instead of low enough for step eight. I did not notice this until after data collection. In the post-data, it was clear that after completing all three tasks that she understood that a smaller *p*-value provides evidence to conclude the alternative hypothesis. Therefore, this could have been a mistake or her understanding regarding *p*-value was still changing.

In conclusion, Carrie was an interesting case for many reasons. Carrie was not prepared to teach statistics and lacked many of the CCK elements that are critical to effectively teach this subject. However, by engaging in only three simulation tasks, Carrie was able to develop her CCK for hypothesis testing, especially regarding the KDUs of the logic of hypothesis testing and the probabilistic nature of hypothesis testing. The focus of the simulation approach on concepts and logic versus procedures and the visualization aspect of the tasks seemed to help develop her understanding. Finally, Carrie envisioned the steps of a simulation approach similar to Lane-Getaz and Zieffler (2006) but added additional steps that seemed to be influenced by the lesson plan design. In the next section, I will present the results from my second case, Kathleen.

Kathleen

The second case, Kathleen, seemed at ease describing hypothesis testing in the pre-interview. I believe this may have been partially from her age. She was over 20 years

older than both other participants. Additionally, Kathleen had the most experience teaching statistics. However, although she has taught statistics for 8 years, she had only been teaching hypothesis testing for the last two years, when she began teaching it as a dual credit class. The students enrolled in dual credit statistics must take a test at the end of the year, like AP statistics, and if they score high enough they receive college credit.

Regarding hypothesis testing, she said, "To be honest, I am not comfortable enough with it either. This will only be my third year to teach hypothesis testing. I'm not to that point where I am really super comfortable. So, that's part of the problem too" (Pre-interview, October 19th, 2018).

Kathleen said that she began teaching statistics because no one else wanted to and when she tells her colleagues that she is teaching hypothesis testing that they complain how difficult it is. When talking about how her students feel about hypothesis testing, she said, "What's interesting to me is that some of my top students can really struggle with this. That's just incredible to me" (Pre-interview October 19th, 2018).

Kathleen's journey with simulations is also enlightening. Although her initial CCK for hypothesis testing was stronger than Carrie's, the simulation tasks were able to develop and deepen several aspects of her understanding. I will begin by presenting the data that is used to answer research question number one, "How does engaging in simulation tasks for hypothesis testing influence high school statistics teachers' understanding of traditional hypothesis testing?" As I did with Carrie, I will share what the data revealed about Kathleen's understanding of hypothesis testing before engaging in the simulation tasks and then compare that with her understanding of hypothesis testing after engaging in the tasks.

Research Question One

Pre-data. Just like Carrie, I had Kathleen complete a pre-open-ended response survey (see Appendix B), pre-CAOS test (see Appendix C), and a pre-interview (see Appendix D). I assessed the data to determine if she possessed the CCK for hypothesis testing by analyzing the pieces of data corresponding to the five KDUs of the logic of hypothesis testing, probabilistic nature of hypothesis testing, importance of data collection in hypothesis testing, importance of sampling distribution in hypothesis testing, and importance of variability in hypothesis testing. The descriptive narrative for this section was produced from the synthesis of the notes produced from using the analytical framework. I organized this section like I did for Carrie and will discuss each of these KDUs in order.

Logic of hypothesis testing. The data revealed that Kathleen possessed a basic understanding of the CCK components for the logic of hypothesis testing that are listed in the analytical framework. However, as will be shown, she could not clearly articulate all her ideas in this area. When describing a hypothesis test, Kathleen wrote,

> When someone makes a claim, we don't necessarily have to decide whether it is true or false. We can state the claim and define an alternate claim. Then we use statistics to evaluate the claim and either let it stand or reject it (Pre-Open-Ended Response Survey, October 5th, 2017).

This showed that she understood that two competing hypotheses are needed. This response also revealed that she understood that failing to reject the null hypothesis does not prove the null hypothesis. The following excerpt also highlighted her understanding of this concept.

Researcher: If we don't reject the null hypothesis what does that tell us or what does it indicate about the truth of the null hypothesis?

Kathleen: If we fail to reject, that really means we're just sticking to status quo. It doesn't mean that we accept it, but we didn't fail to reject it.

Researcher: So, does that mean it's true then?

Kathleen: I teach mine that it doesn't, but I've read somewhere that it means that it's true. But I teach them that it doesn't. Even on the dual-credit test it said to accept it, but that's not Ok.

Researcher: So how do you explain to them why it doesn't make it true?

Kathleen: I go back to the jury. It's guilty or not guilty. They never say they're innocent.

Additionally, she was the only participant that was able to correctly work both a one-sample *z* test for proportions and a two-sample *t*-test for means without consulting any sources. She correctly identified both the null and alternative for each problem. Also, she was able to determine when the test was one-sided and two-sided for both problems, showing that she understood that the practical needs of the researcher should determine if the test should be one or two-sided. She also correctly indicated that statistical significance does not mean practical significance on the pre-CAOS test.

However, Kathleen did show some confusion regarding the logic of a hypothesis test in the context of a simulation. I showed her the same problem that I did Carrie, which asked if she believed the claim that people prefer Pepsi over Coca Cola based on evidence from a sample in which 60% preferred Pepsi. I provided her with the same simulated sampling distribution of sample proportions, which assumed that the

population was split 50-50 in their soda preference. In the simulation, a proportion of 60% or higher occurred only 2.5% of the time. I asked Kathleen to think aloud about the problem. Here is her response.

> But then when I look at this, I think if we took the thought of 50%, the chance to get 60% or higher is only 2.5% of the time. So, it would be really hard to argue. It's almost like this one was a fluke. Like it can happen. So, if we were using an alpha of .05, and this was our hypothesis. This is smaller. So, we would reject. So, I would say that probably I would disagree with that. So, I would argue against that there are more people in the sampled population that prefer Pepsi. Hmm interesting. (Pre-interview October 19th, 2018)

Kathleen stated that you should reject, but she concluded what should be the null hypothesis of there is not a preference for Pepsi. She seemed at first to show the same instinct as Carrie to commit to the null hypothesis. Even though the sample was rare, she thought of it as a fluke and not as evidence to conclude that more favored Pepsi. However, when I attempted to ask her my next interview question, she was reluctant to move on and was unsure about her previous answer regarding the Pepsi problem.

Kathleen: You said it was 60? Wait, I'm still thinking about that problem.

Researcher: Do you want to look at it some more? (I show her the question and simulated sampling distribution again).

Kathleen: Yes, this bothers me, and I still have to sit here and think. Because I haven't thought about this in a year. So, I'm looking at 60% (pointing to where this is located on the simulated sampling distribution). So, I'm going to stick with it.

She stayed with her incorrect response, until a couple of questions later when I asked her to explain when to reject.

Researcher: So how do you explain why we reject when there's a small *p*-value?

Kathleen: So, I always say to them when the *p*-value is that small, you reject because it's not just a fluke that it happened. But it couldn't have happened just by chance. The probability of it happening is so small that the fact that it did happen there's some truth to it. So that's probably what I should have said on the Pepsi one, but anyways, it is what it is. So, we spend some time talking about that. Like if I've got a 30% chance that this could happen, then it can happen sometimes. Don't be shocked about it. If we count m&ms and we do all sorts of things to kind of test that. But I give them little packages of M&M's, and we count them and what's the chance that someone opens their fun size bag of M&Ms and got all blues? Then it's not a fluke, and somebody did that. (She seemed to realize that she made the incorrect conclusion for the Pepsi example and started looking back at the simulated sampling distribution).

Researcher: So, did you say you want to add something to how you explained that?

Kathleen: Yes, because it's the fact that it has such a small chance that it's not a fluke. So yeah, 60% of the people did. There is a preference. Yeah, I just corrected myself!

Upon reflection, I should have asked her to clarify her thinking process more, but after she corrected herself, I went on to the next question. Without a follow-up question, I can only infer the reason she changed her mind. She was explaining the reasoning behind why a small *p*-value led to rejecting the null hypothesis and that seemed to emphasize to her that the Pepsi sample was surprising enough to reject that only 50% favored Pepsi. She was able to correct her answer by focusing on the logic behind a hypothesis test. The next KDU after the logic of a hypothesis test is probabilistic nature, which will be discussed next.

Probabilistic nature. The first component of probabilistic nature is understanding that a cut-point is necessary when determining whether to reject or fail to reject the null hypothesis. Kathleen correctly compared her *p*-value to an alpha-level on the pre-open-ended response survey. She also correctly explained in her pre-interview that you reject when the *p*-value is less than alpha. She emphasized that students still struggle with this concept. She said, "Yes, they get it backwards and we do, I do preach that if your *p*-value is smaller you reject" (Pre-interview October 19th, 2018).

However, Kathleen seemed to struggle with understanding that the alpha-level was the probability of committing a Type I error. She missed the related question on the pre-CAOS test and was unable to explain her reasoning in the interview. The following excerpt from the interview shows her confusion.

Researcher: So, what about type 1 and type 2 errors? Can we actually calculate the probability of those?

Kathleen: Yes, but I always have to look it up. I know Type I is the easy part. One of them is one minus the *p*-value, and the other one is hard to calculate. But we can, but I don't know if we did it in our book. I think they gave it to us. No, it actually says you would learn it in a second stats class. I just know one is hard and one is easy.

The probability of committing a Type I error is the easiest to calculate, because it is equal to the alpha-level. However, it is not calculated as 1 – *p*-value, as Kathleen described one of the errors could be calculated. To calculate the probability of a Type II error, a specific alternative, not just less than, greater than, or not equal to, must be quantified. The probability of a Type 2 error can also be determined by 1 – Power, but not 1 – *p*-value.

Although she was unsure about the relationship between alpha and Type I error, she did seem to understand that changing the alpha-level did have practical consequences. When I asked her, what was the point of setting a significance level, she said,

> Well .05 is pretty typical. You can use .1 or you can use .01. And the one thing I talked to them a little bit about is if it was about some potato chips, I might go with a .1. But if it's about a new cancer drug, then I might want .01. So, if that significance level is .01, you're pretty sure. It makes it more certain in my mind, and so if they're going to validate that new cancer drug, I want them to be positive. And so whatever hypothesis testing they've done, I want it to be as

> accurate as possible. But the bag of potato chips, .1 or whatever. (Pre-interview October 19th, 2018).

This showed that she understood that stronger evidence was required if your alpha-level was smaller.

Kathleen was able to correctly answer all questions regarding *p*-value in the pre-CAOS test and pre-open-ended response survey. She knew the definition of *p*-value, could interpret it correctly, and knew when to reject. For example, on the pre-open-ended response survey (October 5th, 2017) when asked to interpret the *p*-value, she wrote, "Assuming that 67% of the authors support continuing this system, there is a 31.4% chance of obtaining a proportion as large as 72/104 or larger" (October 5th, 2017). She also correctly failed to reject the null hypothesis for this problem and concluded that she could have made a Type II error. The KDU for data collection will be discussed next.

***Data collection*.** Kathleen correctly recognized the presence of bias in the sampling technique in question 8 of the CAOS Tests of Significance. The question stated that a student wished to conduct a significance test but that the data was obtained from people mailing in letters, not a random sample. Kathleen correctly indicated that a test should not be conducted because the conditions were not met.

She also recognized that the way the data is collected influences the type of inference that can be drawn. For the Pepsi problem, she noted,

> So, the first thing that comes to my mind, with just one sample is, the first thing I would think of is, this does say that 60% of this sample favor Pepsi, but can we apply that to a bigger population? So, I know that I am probably supposed to tell

> my students that no you can't do that because these people were not collected from all over the world. (Pre-interview October 19th, 2018)

She knew the importance of how the sample was obtained and recognized the importance of sample size. This was also shown in her written response to pre-open-ended response question one that asked her if she believed a claim of the administration that 80% of their graduates were satisfied, when a sample of size 20 found only 60% were satisfied. She wrote, "Yes. If the survey is conducted by other students, there may be response bias. One survey is not enough (especially with 20 responses) to discredit the claim" (Pre-Open-Ended Response Survey, October 5th, 2017). Although her response shows both an appreciation for the sampling technique and an acknowledgement that sample size is important, it also indicates that Kathleen does not appreciate that a hypothesis test can be used to draw a conclusion, even when the sample size is smaller. The sampling distribution takes the sample size into consideration when showing the variability that one can expect.

Also, she was unable to articulate clearly what types of inferences can be drawn depending on the type of randomization used.

> Researcher: Does it matter for interpreting results if our data comes from a random sample or randomized experiment?
>
> Kathleen: That's a good question. I would think that there's probably a difference when you interpret, but I don't know what I would say different. But it would have to be interpreted differently because in an experiment you have a control group, so there has to be a different interpretation, but I am not knowing what that is.

However, although she could not describe the differences regarding the type of inference that can be made between a random sample and random experiment when asked about them explicitly, her responses in her pre-opened-ended response survey and pre-interview indicated that she knew when a random sample was used one is trying to draw inferences about a population and that random assignment was used when one is trying to establish cause and effect. On the pre-open-ended response survey, question four had a random sample, and question five had random assignment. Her conclusion for both showed the correct type of inference that should be drawn. The last two KDUs for hypothesis testing will be discussed next.

Sampling distribution and variability. The final two KDUs for hypothesis testing, sampling distribution and variability, are reported together in this section. Kathleen's pre-CAOS test indicated that she did understand these topics. She did not miss any of the questions regarding these topics. In the pre-interview, I tried to have Kathleen explain more about them.

Researcher: Can you explain the difference between the population, a sample, and a sampling distribution, and also talk about how their variability might be different?

Kathleen: So, population is everything. So, if we're doing a sample within the school. If we're wanting to do a sample of people that eat out, then the population would be all people that eat out. So, the population are all the units, and the sample . . . it's just a small part pulled from that. So, the variability, if you sample everybody you get what you get, but the variability of the sample, it just depends

> on how random it is. But if I sample one person from every homeroom then I could get the smartest and the weakest for every homeroom. So, there's a lot of variability, but the sampling distribution is all the possible samples. So, in the sampling distribution that variability is going to be constant every time. The samples are going to vary from each other, but when you put it all together in one sampling distribution you get one variability.

Like Carrie, Kathleen was not able to clearly articulate these ideas, and she also did not explicitly state that the sampling distribution was the collection of statistics calculated from the sample. However, she did show that she understood that there are three distinct levels of population, sample, sampling distribution, and that there is a different variability associated with each.

In conclusion, the pre-data indicated that Kathleen possessed a strong understanding of most of the foundational topics of hypothesis testing. She could correctly carry out the steps of a hypothesis test problem, understood why these steps were important, and interpreted a *p*-value correctly. She also understood the differences among population, sample, and sampling distribution and recognized that samples vary. She did struggle some with the logic of hypothesis testing by not recognizing that a hypothesis test can be used to draw conclusions even when the sample size is only 20. Additionally, she initially showed a reluctance to reject the null hypothesis even with a small *p*-value from a simulation. However, she was able to correct herself when she related the simulation to traditional hypothesis testing. Additionally, Kathleen did not show a deep understanding of alpha, the significance level. She knew to compare the *p*-

value to this value to reject or fail to reject the null hypothesis, but she did not associate this number with the probability of a Type I error. In the next section, the post-data will reveal the changes in Kathleen's CCK for hypothesis testing that occurred after engaging in the simulation tasks.

Post-data. After completing all three tasks, Kathleen completed the post-CAOS, post-open-ended question survey, and post interview. This post-data narrative was also produced from the synthesis performed after using my analytical framework for CCK to analyze the data and determine changes in Kathleen's understanding of hypothesis after the simulation tasks. I have organized this section based on the five KDUs of CCK from the analytical framework, like I did for the pre-data.

Logic of hypothesis testing. The pre-data indicated that Kathleen possessed a strong understanding of this KDU, but she had to think carefully about the components to answer the questions correctly. For example, even though she was the only participant to correctly interpret the simulation Pepsi problem, she did so only after making a mistake first. She showed a commitment to the null initially but then changed her answer after discussing the *p*-value. However, in the post-interview (December 19th, 2017), she showed no hesitancy about interpreting the simulation results of the Pepsi problem as shown in the dialogue below.

Researcher: We want to use this information to argue for or against that there are more people from the sampled population who prefer Pepsi than prefer Coca-Cola. And this is simulating that it's not a preference and simulated a whole bunch of times. And so, with this

particular one we've got the 60 percent here at the end. That says .025.

Kathleen: Right. So that tells me that the likelihood of it just happening by chance is very unlikely. So, the fact that they collected a sample where 60 percent favored Pepsi, I would have to go with.

Researcher: They would favor Pepsi? Are you going to argue that?

Kathleen: 60 percent preferred Pepsi. So, I would have to believe that more people prefer Pepsi.

In comparison to the pre-interview, she did not hesitate to make the correct conclusion and confidently stated her answer. This showed that she understood that the simulation was based on the null hypothesis being true and that the fact that the sample was unlikely should lead to rejecting the null hypothesis.

Probabilistic nature. For the pre-data, Kathleen showed a strong understanding of this KDU, except regarding the significance- level. However, for the post-data, she correctly answered the CAOS question that she missed previously regarding the significance level. This question asked how many statistically significant results would be expected if the null hypotheses was true and the significance level was 5 percent. Kathleen had previously selected zero, but she corrected her answer to 5 on the post-CAOS. Additionally, the post-interview showed that Kathleen had developed a better understanding of the significance level. When Kathleen conducted her own simulation for the satisfaction rate for the university problem, she did not specify a significance level, so I asked if she had thought of one when making her conclusion.

Researcher: Right. So that will be your conclusion on that. So, did you think of a significant level or did you pick one?

Kathleen: I may be going on a little bit of a tangent, but the significance level never meant anything to me, and it does now. OK, it does now, because it's sort of like a guideline. Like if this was somebody's project, and you had a project manager and said be sure you use a significant level of .05. That tells me if it's less than .05, I've got something significant going on here. And if it's more than that then yeah no. There's your guideline. Yeah. In reading the textbook that's never made sense. And if it's just sort of iffy to me. I know it doesn't mean anything to my kids. It does now because I'm seeing this number right here.

Kathleen understood that the point of setting a significance level was to determine before-hand how much evidence you were going to require. She said it made more sense because she was able to visualize that rejection region. For the same problem, she was also able to identify when you would reject by corresponding that to the significance level as will be shown in the following excerpt.

Researcher: Which values from the study would make us reject?

Kathleen: So now, 65 percent, 65 or less, and that would. And that makes sense because there's just so many. There's just, even if you didn't look at these numbers, 60 percent to 80 percent, 65 percent to 80 percent is a little more reasonable. Yeah, and so you can kind of

> see that OK there's that boundary. And that's even, you know, way over here so it's nice to see that.

Kathleen recognized that there were many samples between 65 and 80%. Therefore, sample proportions in that range would be reasonable if the approval rating was 80%. However, where the rejection region started, there were very few sample proportions in that area, indicating it was unlikely to get a sample proportion that low if the approval rating was truly 80%. Kathleen was able to visually see the sampling distribution and the alpha-region.

***Data collection*.** Kathleen did not show any deficits in this KDU in the pre-data. She also got all the corresponding questions correct in the post-data. In the post-interview she continued to focus on making sure that bias was not present as one of the most important things to do before conducting a test. Before conducting a simulation for the university satisfaction problem, she said,

> But when somebody just walks up to you, and when is this question asked? Is it after they walk across the stage, and they have that diploma in their hand? Then a whole bunch of them are satisfied. That this is out on the street and after they've taken a midterm or something. Now they may not be satisfied. So that could be a little bias creeping in there. (Post-interview, December 19th, 2017)

The random condition is the most important condition for simulations, and Kathleen consistently thought about the implications of where the data came from. Next, I will discuss sampling distribution and variability.

Sampling distribution and variability. The pre-data indicated that Kathleen possessed a strong understanding of this KDU. Therefore, this area was unaffected by the

simulations. However, Kathleen did think that the power of the simulation was in the ability to see the construction of the sampling distribution. She said, "And when you can watch this build that's just incredible. Yeah, I love that" (Post-interview, December 19th, 2017). Also, she acknowledged that the students would be able to see the variability and determine if a sample was likely or not. She commented, "Yeah. And to be able to watch this and build and do one sample and here's what we got. And then the students when they watch that more and more and more, they're like, oh well, .7. There's a whole bunch though. Yeah that's not surprising" (Post-interview, December 19th, 2017).

In conclusion, Kathleen did possess a strong understanding of hypothesis test initially. However, the simulation tasks helped her deepen her understanding of the logic of a hypothesis test and helped her understand the significance level better. Kathleen focused on the visualization aspect of simulations and felt that was powerful to help students overcome misconceptions. She wrote, "The calculations were made simple due to technology, and the reject/fail to reject that students struggle with vanished! Students are able to see whether a claim could be possible or if it is undeniably inconceivable" (Post-open-ended survey, November 27th, 2017). In the next section, I will discuss the overall themes for Kathleen for each research questions.

Themes. Themes for Kathleen were determined by the inductive analysis and in-depth reflection described in Chapter Three. For Kathleen, two major themes emerged for research question one. The first theme for research question one was that simulations focus on visualizations, which help develop the KDU of probabilistic nature. Second, simulation tasks focus on concepts and logic, which develops the KDU of the logic of hypothesis testing.

Research Q1, Theme One. The first theme was that simulations focus on visualizations, which help develop the KDU of probabilistic nature. From the first task, Kathleen's focus was on the visualization aspect of simulations and connecting this to an understanding of *p*-value and significance level. At the end of the first task, she said, "It's really neat because they can see that. We figured this out, and we didn't talk about *p*-value one time, but they see it" (Task A, November 9th, 2017). She also wrote, "By performing simulations, we were able to see what happens, which in this task, helped us to really understand how significant 14 out of 16 is. It was much easier to draw a conclusion based on what we saw instead of comparing *p*-values to alphas" (Post-Task A Reflection, November 9th, 2017).

This trend continued for task B. She wrote, "During the simulations, students are able to see how the simulations work, and with the simulator technology, they can see literally thousands of data points. Finally, students have an easier time making a conclusion, because they can see the information" (Post-Task B Reflection, November 15th, 2017). It was also during task B that Kathleen recognized that a Type I error could have occurred if we reject the null hypothesis. She did this when looking at the simulated sampling distribution. It was Chase who stated that the probability of this occurring would be 5%, but Kathleen retained the knowledge that your significance level determines the probability of a Type I error as reported in her post-data section. Also, as shown with the included quote in the post-data section, Kathleen stated that she believed it was the visualization aspect of simulations that helped her finally understand the alpha-level.

Kathleen's focus on the visualization component continued in both task C and in her post-interview. She wrote, "After simulations (both small and with technology) we can see the *p*-value. Evaluating the *p*-value is much easier, because it makes so much more sense! We are able to see how rare or often this would happen" (Post-Task C Reflection, November 20th, 2017). In the post-interview she said,

> So, when I'm not having to, and I'm talking in this student voice right now, when I'm not having to really focus on how to get that *p*-value. What is that doing? Am I rejecting or failing to reject, what not? She told me memorize those things. Now I'm talking about *p*-value, if I'm using .05, the administration is just wrong. Yeah, they're just wrong. So, it all, it just makes sense and comes to us and makes it make sense. And when you can watch this build that's just incredible. Yeah, I love that.

Kathleen was expressing how the *p*-value could be seen on the sampling distribution instead of having to calculate it. Although Kathleen could interpret *p*-values correctly in the pre-data, she felt that the *p*-value made more sense now. The visualization component helps to logically and more easily make a conclusion instead of having to complete all the computation steps. I will share the second theme for research question one next.

Research Q1, Theme Two. Simulation tasks focus on concepts and logic instead of procedures, which develops the KDU of the logic of hypothesis testing. The six-phase lesson structure used to develop the tasks in this study were designed based on NCTM's mathematical practices. One of the practices is to facilitate meaningful mathematical discourse (NCTM, 2014). Thus, a critical element of the tasks was to promote discussions

that would help students focus on the underlying concepts and logic of a hypothesis test. Kathleen wrote,

> The simulation (with discussion) leads to a better understanding of the hypothesis (instead of Ho; Ha). There is more discussion in the task as we made a guess. We were able to reason through our guesses and consider if we're guessing 10, that means only 6 chose the bad guy. (Post-Task A Reflection, November 9th, 2017)

This component of discussing the reasoning behind each hypothesis is not part of a simulation approach as discussed in the literature in Chapter Two. This incorporation was completely determined by the lesson plan. Kathleen's statement expresses the idea that the simulations focus on reasoning through the test instead of simply following the steps of a hypothesis test.

Kathleen continued this idea of focusing on the concepts and logic versus procedures, when she wrote the following after task B,

> In the traditional approach, the information is stated, hypotheses are written, and calculations are performed. In the simulation approach, students begin by making guesses and evaluating optional hypotheses. Before any calculations are done, students think about values that would be surprising and what is expected. (Post-Task B Reflection, November 15th, 2017)

Kathleen describes the traditional approach as a list of steps. In contrast, Kathleen says that in a simulation approach, students begin by making guesses and evaluating hypotheses, which focuses on the concept of determining competing hypotheses. Additionally, she stated that students think about what would be surprising and what would be expected, which focuses on the concept of variability.

Evidence that the simulation tasks used in this study focus on the logic of a hypothesis test was also seen in the post-open-ended-response survey. Kathleen wrote, "The way to introduce hypothesis testing to students is now clear. The focus is not on memorizing reject/ fail to reject rubrics; rather, students can now logically think through a test" (November 20th, 2017). With this statement, Kathleen is acknowledging the importance of having students logically think through a test instead of memorizing the steps and procedures. By doing so, Kathleen's own development of understanding the logic of a hypothesis test was seen when she worked the Pepsi problem again in the post-interview. She did not have to hesitate or think through what the conclusion would be. In the pre-data she eventually was able to correct herself, but only after discussing that a small *p*-value was when you should reject the null hypothesis. She was able to use the rule of rejecting when the *p*-value was less than alpha, but she was not focusing on the logic of why a small *p*-value meant to reject. However, when she worked this problem in the post-interview, she readily understood that the sample was being used as evidence to assess whether most people favor Pepsi from the sample population. She recognized that the small *p*-value meant the sample data was surprising enough to conclude that more people favor Pepsi. She did not have to base her conclusion on the memorized step that a *p*-value smaller than the alpha-level means to reject the null hypothesis.

Also, in the post-interview, I had Kathleen work the problem concerning university satisfaction using simulations. This problem asked one to evaluate the claim of a university that over 80% are satisfied, based on a sample in which only 60% were satisfied. She focused on the logic of the test when explaining how she would begin to work the problem with her students. She said, "So then I would be talking about how do

we set it up? So, I don't like all that dialogue at first is missing, because I feel like in the textbook that's what you don't get" (Post-interview, December 19th, 2017). Kathleen was expressing that in traditional textbook problems she believes that students are simply following a procedure of stating the null and alternative. By adding the component of discussing what you are logically trying to do in this scenario, students can think about their reasons for selecting the different hypotheses. Then at the end of the problem, she said,

> I don't think you've asked me to reject or fail reject not one time. I just made a conclusion. Yeah, because so it's, if they go back to the problem and say we think administration is wrong. There's your answer. They're not right in their claim. That's all that we're asking them to do. (Post interview, December 19th, 2017)

Kathleen was referencing how in traditional hypothesis testing the focus is on following the steps and arriving at the prescribed reject or fail to reject instead of simply focusing on what question you are trying to answer. The heart of the test is not about rejecting or failing to reject but thinking about if you believe a claim or not based on the evidence.

In conclusion, the data in this section answers research question one by sharing what aspects of Kathleen's CCK changed after engaging in the simulation tasks as determined by comparing Kathleen's evidence of understanding with the analytical framework. Kathleen gained a deeper understanding of the logic behind a hypothesis test and the probabilistic nature of hypothesis testing. Additionally, the inductive analysis of the data was shared, which showed the influence of the simulations' focus on concepts and logic and the visualization aspect of simulations. In the next section, the data which

answered research question two will be shared, along with the theme for research question two.

Research Question Two

Research question two was, “How do simulation tasks influence high school statistics teachers’ understanding of simulations and how do they make connections between traditional and simulation approaches for hypothesis testing?” To answer this research question, I will share the results of the deductive analysis used to analyze the post-task reflection data. The results include a descriptive narrative of the post-task reflections and the tables used to compare Kathleen’s simulation models and connecting approaches. Additionally, I will share the results from the inductive analysis that were used to determine the theme corresponding to research question two.

As part of the post-task A reflection, I had Kathleen draw a diagram representing the simulation from the task. Figure 12 shows her diagram.

1. Draw a diagram representative of the simulation task that you just completed.

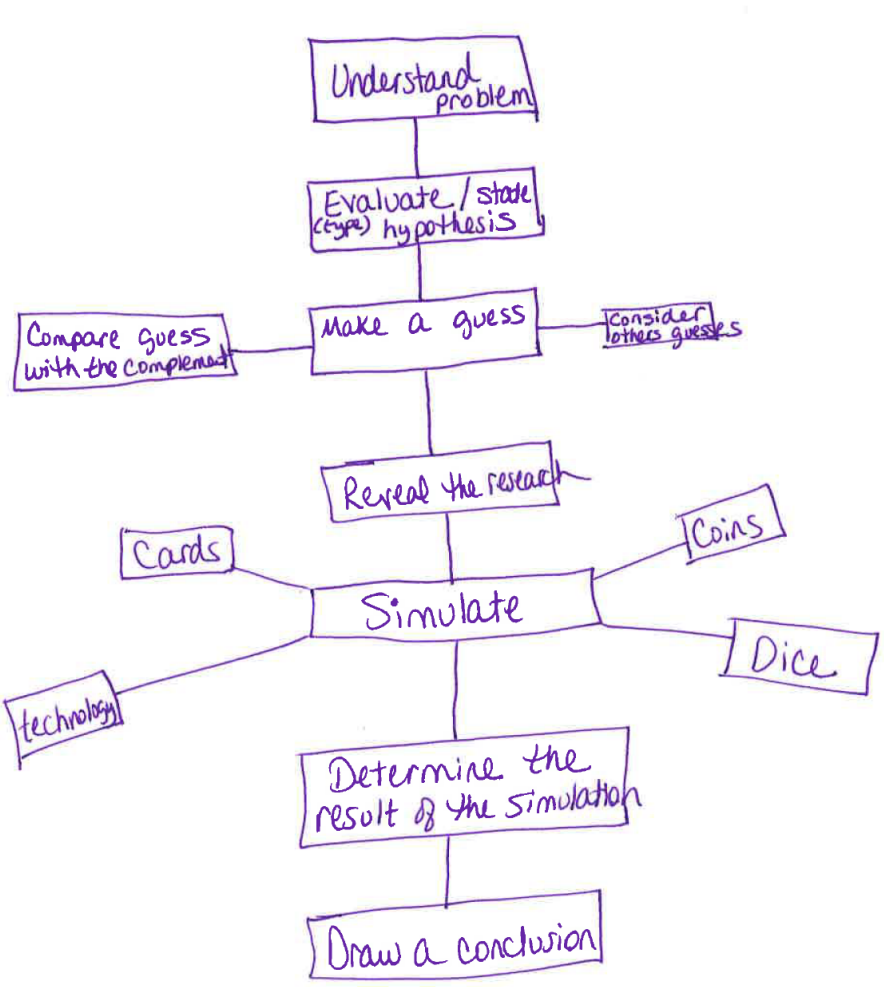


Figure 12. Kathleen's Simulation Model, Post Task A Reflection (November 9th, 2017)

The following table (see Table 19) shows how her diagram corresponded to Lane-Getaz and Zieffler's (2006) modified SPM described in Chapter Two. Like Carrie, Kathleen added additional steps of making a guess and revealing the research, which seemed to be influenced by the lesson plan design. Also, Kathleen split Lane-Getaz and

Zieffler's (2006) step one and three into two steps each. However, the overall meaning was the same.

Table 19

Alignment of Lane-Getaz and Zieffler (2006) Modified SPM with Kathleen's Post-Task A Model

Lane-Getaz and Zieffler (2006) SPM	Kathleen's Model
1. Establish population parameters	1. Understand the problem 2. Evaluate/State hypotheses
	3. Make a guess
	4. Reveal the research
2. Generate samples through simulation.	5. Simulate
3. Create sampling distribution and assess unusualness.	6. Determine the results of the simulation 7. Draw a conclusion

Next, Kathleen was asked to work a traditional hypothesis problem that correlated to the task. The following figure (see Figure 13) shows her response.

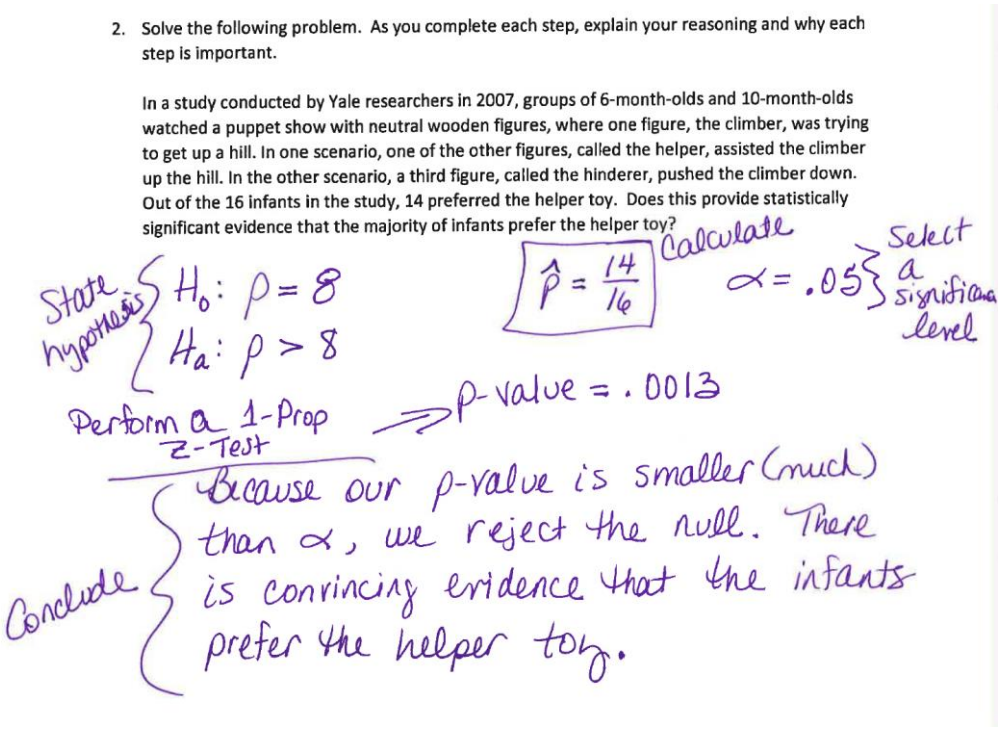


Figure 13. Kathleen's Post-Task A, Question Two Response

She stated a correct pair of hypotheses and indicated the calculator command that she used to solve the problem. However, she did not report the *z*-test statistic or check conditions. Her conclusion was correct. She was also asked to connect her simulation model that she drew to the traditional hypothesis test problem that she worked. Like Carrie, she did not explicitly link the steps. Instead she discussed the simulation task in general. She wrote five bulleted points,

- The simulation (with discussion) leads to a better understanding of the hypothesis (instead of Ho; Ha).

- There is much more discussion in the task as we made a guess. We were able to reason through our guesses and consider "if we're guessing 10, that means only 6 chose the bad guy"
- By revealing the research, it made the information much more interesting. It was revealed after discussion, instead of in the beginning.
- By performing simulations, we were able to "see" what happens, which in this task, helped us to really understand how significant 14 out of 16 was.
- It was much easier to draw a conclusion based on what we "saw" instead of comparing *p*-values to alpha. (Post Task A Reflection, November 9th, 2017).

Her points illustrated how she believes the simulation approach provided a more natural and in-depth way to develop the population parameters. Instead of writing the null and alternative as the initial step, the lesson plan asks the participants to make a guess about what they believe the study's results will be. Additionally, instead of focusing on traditional notation to write the null and alternative, the lesson plan had the participants discuss all the different possible hypotheses and write them out in words in the context of the problem. Participants are then told which one of their hypotheses is the null and asked to discuss what would be the most likely outcome if the null was true and what other types of results would not be surprising. All of this is done before the actual study's results are revealed. This provides a way for participants to really understand what the population parameters are that they will be simulating. Also, Kathleen focused on how the simulations allowed them to "see" what happens instead of using the traditional method of comparing the *p*-value to alpha.

After the second simulation task, I asked Kathleen to create a step by step guide that students could use to conduct a hypothesis test using simulations. I asked her to list steps instead of the diagram to see if I could obtain more details regarding her thinking about the simulation design. I again allowed her to have the task handout to reference as she completed this. The following shows how Kathleen listed the steps.

1. Read the task and make a guess regarding thoughts on the outcome.
2. Determine the hypotheses that can be made.
3. Make a statement regarding an outcome that would surprise. Create a boundary.
4. Share the results of the actual experiment.
5. Design a simulation: roll dice, draw from a lot, random number generator, Table A? State precisely this design.
6. Carry out the simulation and record results.
7. Discuss results (summarize).
8. Simulate with technology and compare results.
9. State the *p*-value and write a conclusion.

These steps were like the steps shown in the diagram she created representing the simulation after task A. However, she added discussing results and listed simulation by technology separately in her steps after task B. These steps again seem to be influenced by the lesson plan. Therefore, I have provided a table (see Table 20) showing how Kathleen's model corresponds to Lane-Getaz and Zieffler's (2006) modified SPM with the pedagogical steps that do not represent a simulation omitted.

Table 20

Alignment of Lane-Getaz and Zieffler's (2006) Modified SPM with Kathleen's Post-Task B Model

Lane-Getaz and Zieffler (2006) SPM	Kathleen's Model
1. Establish population parameters	2. Determine the hypotheses that can be made.
2. Generate samples through simulation.	7. Simulate with technology and compare results.
3. Create sampling distribution and assess unusualness.	9. State the *p*-value and write a conclusion.

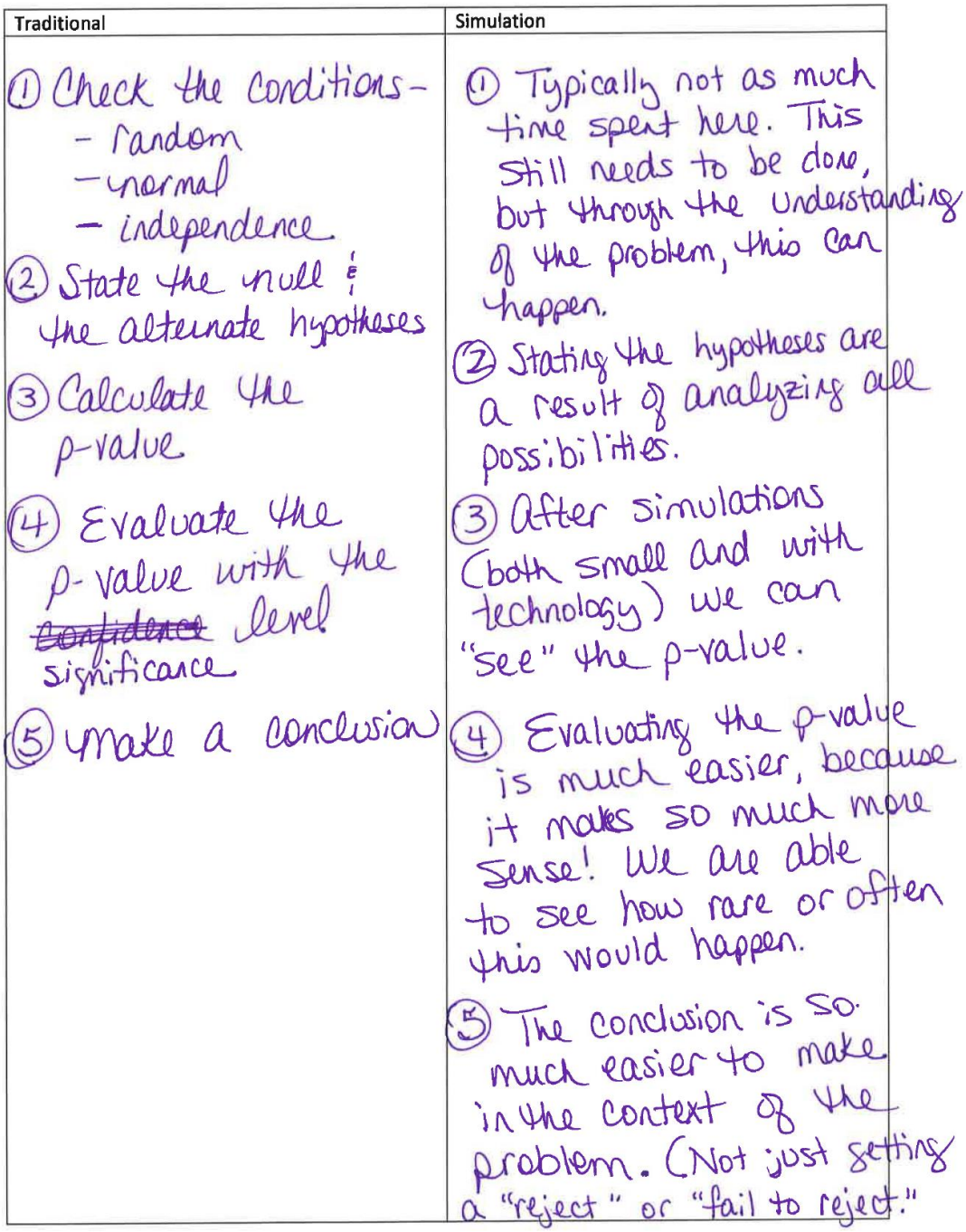
Like after task A, her steps align with Lane-Getaz and Zieffler's (2006) modified model from the analytical framework, except the phrases she used were slightly different, and she added additional steps based on the lesson plan. Also, after the task, I asked Kathleen to explain how she saw each step of the traditional hypothesis test connected to the simulation approach. She wrote,

> In the traditional approach, the information is stated, hypotheses are written, and calculations are performed. In the simulation approach, students begin by making guesses and evaluating optional hypotheses. Before any calculations are done, students think about values that would be surprising and what is expected. During the simulations, students are able to "see" how the simulations work and with the simulator technology, they can see literally thousands of data points. Finally,

> students have an easier time making a conclusion, because they can see the information.

She described the differences between the two approaches, but she did not explicitly show how the steps were connected. Her explanation does show that she believes that the traditional approach focuses on procedures, and simulations, on the other hand, allow students to visualize beyond the calculations.

After the final task, I provided a chart with traditional listed on the left and simulation listed on the right. I did not have Kathleen list the steps to a simulation approach separately, like I did in the previous two tasks, so that I could provide Kathleen with a way to more explicitly connect the two approaches. I asked Kathleen to list the steps needed for each approach and to line them up. I emphasized to not think about how to do this according to a lesson but to focus on the essential steps needed to solve a problem using each method. The following figure (see Figure 14) shows Kathleen's diagram.



Traditional	Simulation
① Check the conditions – - random – normal – independence ② State the null & the alternate hypotheses ③ Calculate the p-value ④ Evaluate the p-value with the ~~confidence~~ level significance ⑤ Make a conclusion	① Typically not as much time spent here. This still needs to be done, but through the understanding of the problem, this can happen. ② Stating the hypotheses are a result of analyzing all possibilities. ③ After simulations (both small and with technology) we can "see" the p-value. ④ Evaluating the p-value is much easier, because it makes so much more sense! We are able to see how rare or often this would happen. ⑤ The conclusion is so much easier to make in the context of the problem. (Not just setting a "reject" or "fail to reject."

Figure 14. Kathleen's Connection of Simulation and Traditional Approach Post-Task C

First, I analyzed the right-hand side of her diagram with the analytical framework to see how her simulation steps aligned with Lane-Getaz and Zieffer's (2006) modified three-tier model. I left out Kathleen's step one of the simulation because it was about how checking conditions from the traditional approach was connected to the simulation approach. She wrote "typically not as much time spent here" (Post- Task C Reflection, November 20th, 2017), indicating that she did not view this step as part of the simulation. The following table (see Table 21) shows how they are connected. To avoid confusion, I have omitted the numbering.

Table 21

Alignment of Lane-Getaz and Zieffler's (2006) Modified SPM with Kathleen's Post-Task C Model

Lane-Getaz and Zieffler (2006) SPM	Kathleen's Model
Establish population parameters	Stating the hypotheses are a result of analyzing all possibilities.
Generate samples through simulation	Not listed
Create sampling distribution and assess unusualness.	After simulations (both small and with technology) we can "see" the *p*-value. Evaluating the *p*-value is much easier, because it makes so much more sense! We are able to see how rare or often this would happen The conclusion is so much easier to make in the context of the problem. (Not just getting a "reject" or "fail to reject"?

How Kathleen listed the steps of the simulation approach was slightly different than the steps she listed after the first two tasks. In the first two tasks, I asked Kathleen to list the simulation steps first, then to work out a traditional test problem. It was after doing both separately, that I asked her to connect the two approaches. However, because she was not explicitly linking the two steps, I had her list the simulation and traditional steps at the same time on post-task C reflection. Therefore, some of her steps are expressed in terms of a comparison. For example, she stated that evaluating the *p*-value is easier for step four, which is obviously in comparison to her traditional step number four of evaluating the *p*-value with an alpha-level.

Next, I used her diagram (see Figure 14) and compared it with the analytical framework to see how her connections compared to the connecting approaches model of Lane-Getaz and Zieffler (2006), as shown in the following table (see Table 22). Note, when connecting approaches, Lane-Getaz and Zieffler (2006) described the simulation approach as five steps in comparison to their three-tier SPM.

Table 22

Alignment of Lane-Getaz and Zieffler's (2006) Connecting Approaches with Kathleen's

Traditional Approach		**Simulation Approach**	
Lane-Getaz and Zieffler (2006)	Kathleen	Lane-Getaz and Zieffler (2006)	Kathleen
Statement of null and alternative hypothesis	State the null and alternative hypothesis.	What if scenario? Determine a model.	Stating the hypotheses are a result of analyzing all possibilities.
Check conditions	Check the conditions (random, normal, independent).	Repeat the random sampling or random assignment through simulation.	Typically, not as much time spent here. This still needs to be done but through the understanding of the problem, this can happen.
Calculate *z* or *t* test statistic	Not listed.	Select the appropriate summary statistic.	Not listed
Graph a theoretical sampling distribution based on a function	Not listed.	Compile summary statistic for distribution formed by simulation.	Not listed.
Assess rareness by large or small *p*-value OR large of small test statistic	Calculate the *p*-value	Assess rareness by finding observed data on simulated sampling distribution and calculating approximate *p*-value.	After simulations (both small and with technology) we can "see" the *p*-value.
	Evaluate the *p*-value with the significance level.		Evaluating the *p*-value is much easier, because it makes so much more sense! We are able to see how rare or often this would happen
	Make a conclusion		The conclusion is so much easier to make in the context of the problem. (Not just getting a "reject" or "fail to reject"?

Kathleen did not include calculating a *z* or *t* test statistic or graphing a theoretical sampling distribution for the traditional approach. The other steps were similar, but Kathleen used more steps when discussing the *p*-value and making a conclusion. Next, I will share the theme for research question two.

Research Q2, Theme One. The theme for research question two was that understanding of simulations and the connection of simulation and traditional approaches was influenced by the lesson plan design. Like Carrie, Kathleen's description of a simulation approach was directly influenced by the lesson plan design. The following table (see Table 23) shows how her simulation steps, which she wrote after completing task B, corresponded to the lesson plan. I used her steps from task B because they were like the simulation steps listed in the other two post-task reflections but provided the most details. The only part of the lesson plan not included was 3a, which asked to state the most likely result if the null hypothesis was true. Also note that some of the numbering is different starting with step five. This is because Kathleen used two steps in place of question five from the task B lesson plan handout.

Table 23

Kathleen's Steps Connected to Task B Lesson Plan

Kathleen's Steps	Lesson Plan
1. Read the task and make a guess regarding thoughts on the outcome.	1. If 20 students took an exam, 10 on white paper and 10 on yellow paper, how do you think the average score from students who took the exam on yellow paper would compare to average score of the students who took the exam on white paper? Why?
2. Determine the hypotheses that can be made.	2. How many different hypotheses could we make for this situation regarding the averages of scores of students who take the exam on yellow paper and on white paper? What are they?
3. Make a statement regarding an outcome that would surprise. Create a boundary.	3b. Still assuming that the color of the exam did not affect students' scores (i.e. students would get the same score regardless of the color of the exam) what kind of results (difference in the average scores) would you not be surprised to see when this study is conducted with 20 participants?
4. Give the actual difference from the experiment.	4. For this experiment, the average test score for the yellow paper was _____ and the average test score for the white paper was _____. Therefore, the actual difference in the average scores of students who took the exam on yellow paper compared to students who took the exam on white paper was _____. If it is REALLY the case that the color of the exam doesn't matter, do you find the teacher's result surprising? Why or why not?
5. Design a simulation: roll dice, draw forma lot, random number generator, Table A? State precisely this design.	5. Determine a plan for simulating the study.
6. Carry out the simulation and record results.	5. Describe your plan below and carry out three trials of your simulation.
7. Discuss results (summarize)	6a. From your results, does it seem like the results obtained by the teacher would be surprising? Explain.
8. Simulate with technology and compare results.	6b. Now, we will use technology to simulate this experiment many, many times under the assumption that the null hypothesis is true. Based on this

Kathleen's Steps	Lesson Plan
	simulation how surprising are the actual results of this study? Explain your reasoning
9. State the *p*-value and write a conclusion	6c. Based on the results of the simulation, how likely is a difference of 6.3 or greater? Explain.

In conclusion, Kathleen initially showed a strong understanding of hypothesis testing. However, her CCK was strengthened by engaging in the simulation tasks. She was able to better understand the logic of hypothesis testing and the significance level. Additionally, how she understood simulations and made connections to the traditional approach was directly linked to the lesson plan design. In the next section, I will provide a cross-case analysis between Carrie and Kathleen.

Cross-Case Analysis

Carrie and Kathleen were purposefully selected to represent literal replications, which is like repeating an experiment in quantitative studies (Yin, 2014). I had anticipated that these cases would be similar enough to serve as literal replications, because none of the teachers had been exposed to simulations previously and each teacher was engaged in the same tasks. I also anticipated that simulation tasks would similarly impact these teachers' understanding of hypothesis testing and how they understood simulations and connected simulation and traditional techniques. Of course, there were individual differences, such as teaching experience and content knowledge. In this section I will report on both the similarities and differences found between these two cases.

The pre-data indicated that Kathleen possessed a stronger content knowledge of hypothesis testing. Kathleen could work out traditional one- and two-sample hypothesis test problems without using resources. Carrie could only work out one-sample problems with resources and was unable to work out any two sample problems. On the pre-CAOS test, Kathleen only missed two questions, and Carrie missed eight. Also, Carrie did not have an in-depth understanding of *p*-value or Type I Error. In the pre-interview she argued that a large *p*-value would result in rejecting the null hypothesis and did not acknowledge that a mistake could have been made when rejecting the null hypothesis. However, both participants did show incomplete understanding for the KDUs of the logic of hypothesis testing and probabilistic nature of hypothesis testing.

Regarding engaging in the simulation tasks, Kathleen contributed more to the overall discussions. However, Carrie contributed more on designing the actual simulations. Also, Kathleen focused on the concepts from the very beginning, but Carrie initially focused on procedures associated with using traditional hypothesis testing. By the third task, Carrie had moved away from traditional procedures and wrote her answers using the context of the scenarios. Also, Carrie was more accurate in predicting the types of results that would be expected if the null hypothesis was true. Kathleen consistently provided a much smaller range of values than would be expected. What each participant focused on during the tasks was also slightly different. Kathleen focused more on the visualization component. Seeing the *p*-value and the alpha-level were common elements that were very meaningful for her among all three tasks. Although Carrie eventually moved away from procedures, she did tend to focus more on them, and she was the only one to question the validity of the simulated sampling distribution in comparison to the

theoretical sampling distribution. I will report more on these similarities and differences after comparing their post-data.

Next, I will compare the post-data, which showed the changes in understanding of hypothesis testing, for the participants. The following table (see Table 24) summarizes the categories of understandings of hypothesis testing that were influenced for each participant, as reported individually for each case in this chapter. For both participants the KDUs of logic of hypothesis testing and probabilistic nature were influenced. However, the number of components under each KDU that was affected was different. For the KDU of the logic of hypothesis testing, the component of indirect reasoning was strengthened for both participants. However, for probabilistic nature, Carrie's understanding of Type I error, significance level, and *p*-value were all affected, and for Kathleen only her understanding of the significance level was affected. However, Type I Error and *p*-value were not categories that the pre-data indicated were lack of understandings for Kathleen. Therefore, understandings that were deepened because of the simulation tasks were similar for the categories in which both participants initially showed deficits in understanding.

Table 24

Categories of Understanding Influenced by Simulations

Carrie	Kathleen
KDU: Logic of Hypothesis Testing	**KDU: Logic of Hypothesis Testing**
Indirect Reasoning	Indirect Reasoning
KDU: Probabilistic Nature	**KDU: Probabilistic Nature**
Type I Error	Significance Level
Significance Level	
P-value	

In comparing the overall themes for the participants, I found similar results for both participants. However, the evidence obtained for each participant was different. The following table (see Table 25) shows a comparison for the evidence obtained for each participant for the theme regarding visualization.

Table 25

Comparison of Evidence for Visualization Theme

Theme	Carrie's evidence	Kathleen's evidence
Simulations focus on visualizations, which help develop the KDU of probabilistic nature	" a 'real' normal curve rather than the one in our books" (Post-Task B Reflection, November 15th, 2017) "Use technology to simulate doing this many, many times *see Normal distribution" (Post-Task C Reflection, November 20th, 2017) "The simulation helped me see how the sample dist. Really does start to look Normal!" (Post-Open-Ended Survey, November 27th, 2017). "So, the simulation just gives us a quick and easy way to see the visualization of the distribution and to quickly find the probability of that value or less happening" (Post-interview, December 18th, 2017).	"It's really neat because they can see that. We figured this out, and we didn't talk about *p*-value one time, but they see it" (Task A Transcription, November 9th, 2017) "By performing simulations, we were able to see what happens, which in this task, helped us to really understand how significant 14 out of 16 is. It was much easier to draw a conclusion based on what we saw instead of comparing *p*-values to alphas" (Post-Task A Reflection, November 9th, 2017). "During the simulations, students are able to see how the simulations work, and with the simulator technology, they can see literally thousands of data points. Finally, students have an easier time making a conclusion, because they can see the information" (Post-Task B Reflection, November 15th, 2017). "After simulations (both small and with technology) we can see the *p*-value. Evaluating the *p*-value is much easier, because it makes so much more sense! We are able to see how rare or often this would happen" (Post-Task C Reflection, November 20th, 2017). So, when I'm not having to, and I'm talking in this student voice right now, when I'm not having to really focus on how to get that *p*-

Theme	Carrie's evidence	Kathleen's evidence
		value. What is that doing? Am I rejecting or failing to reject, what not? She told me memorize those things. Now I'm talking about *p*-value, if I'm using .05, the administration is just wrong. Yeah, they're just wrong. So, it all, it just makes sense and comes to us and makes it make sense. And when you can watch this build that's just incredible. Yeah, I love that. (Post-interview, December 19th, 2017)

For both participants, the power of the simulations was in the ability to see the building of the sampling distribution, which allows one to see the *p*-value instead of having to perform calculations and simplifies the process of making conclusions. This allowed the participants to focus on the meaning of the *p*-value and relate the significance level to an area under the curve. The role of the sampling distribution as a probability model is also made more explicit. However, Kathleen was much more overt about the visualization role and mentioned this component more often. Also, Carrie's evidence focused more on seeing the sampling distribution and just mentions the ability to quickly determine the probability (referencing the *p*-value) once. Kathleen discussed seeing the *p*-value more, and she also focused more on the ability to quickly draw a conclusion when using simulations. Next, I will show a comparison of the evidence for the concepts and logic theme (see Table 26).

Table 26

Comparison of Evidence for Concepts and Logic Theme

Theme	Carrie's evidence	Kathleen's evidence
Simulation tasks focus on concepts and logic instead of procedures, which develops the KDU of the logic of hypothesis testing.	Comparison of Question 2 for each task: Task A: traditional notation only, Task B: mix of traditional with context Task C: only context Pre-data described hypothesis testing as a list of steps. Post-data: focused on the logic behind the test Post-data: Worked out a simulation problem without setting up the null and alternative according to the traditional method but used the logic behind the test to determine the hypotheses.	The simulation (with discussion) leads to a better understanding of the hypothesis (instead of Ho; Ha). There is more discussion in the task as we made a guess. We were able to reason through our guesses and consider if we're guessing 10, that means only 6 chose the bad guy. (Post-Task A Reflection, November 9th, 2017). In the traditional approach, the information is stated, hypotheses are written, and calculations are performed. In the simulation approach, students begin by making guesses and evaluating optional hypotheses. Before any calculations are done, students think about values that would be surprising and what is expected. (Post-Task B Reflection, November 15th, 2017). "The way to introduce hypothesis testing to students is now clear. The focus is not on memorizing reject/ fail to reject rubrics; rather, students can now logically think through a test" (Post-Open-Ended response, November 20th, 2017). "So then I would be talking about how do we set it up? So, I don't like all that dialogue at first is missing, because I feel like in the textbook that's what you don't get" (Post-interview, December 19th, 2017)

Theme	Carrie's evidence	Kathleen's evidence
		I don't think you've asked me to reject or fail reject not one time. I just made a conclusion. Yeah, because so it's, if they go back to the problem and say we think administration is wrong. There's your answer. They're not right in their claim. That's all that we're asking them to do. (Post interview, December 19th, 2017)

Although the same theme of simulation tasks focus on concepts and logic instead of procedures, which develops the KDU of the logic of hypothesis testing was determined for both participants, the evidence used to establish this theme for each participant was different. For Carrie, evidence for this theme was obtained mainly from her work. Her work showed how she was moving away from the traditional steps and procedures of hypothesis testing and began focusing on what she was logically trying to deduce from the evidence. For Kathleen, her evidence was derived mainly from quotes in which she focused on the logic of the test. Additionally, as was the case for the visualization theme, there was also more evidence from Kathleen.

In comparing these results for research question one, simulations developed the KDUs of the logic of hypothesis testing and probabilistic nature for both participants. The KDU of probabilistic nature was influenced by the visualization component for both participants and the KDU of the logic of hypothesis testing was influenced for both participants from the focus of simulations on logic and concepts. However, Carrie showed an inclination to rely and focus on procedures in her pre-data. Her data revealed that she moved away from procedures and focused on concepts through participation in

the simulation tasks. This enabled her to understand the indirect reasoning behind a hypothesis test. However, Kathleen did not show an inclination to focus on procedures. However, the discussion element, which focused on logic and concepts, allowed her to more deeply understand this component of hypothesis testing, as shown in her quotes. For both participants, being able to see the sampling distribution allowed them to visualize the individual components of p-value and the significance level. By doing so, they were able to make sense of these abstract concepts, which they had previously relied on interpreting from the memorization of rules and procedures. Overall, simulations focused on the important concepts of hypothesis testing and offered a way to visualize abstract elements, which deepened the CCK of hypothesis testing for both participants.

For research question two, the same theme was established for both participants. The lesson plan design directly impacted how both participants described the simulation approach. Additionally, how the participants listed the steps of the simulation was very similar. The following table (see Table 27) shows a comparison of how each of the participants' steps are represented in the lesson plan. For each participant, I used the simulation steps that they listed in their post-task B reflections. I used these steps because both participants provided more details in this reflection piece concerning the simulation steps in comparison to the other two post-task reflections. In the table, I listed the numbering used by both participants in their steps and used the actual numbers from the lesson plan task B sheet for reference purposes. Therefore, the numbers are not aligned across the table. For example, Carrie did not list a step which corresponded to the first question on the task handout; therefore, her step one corresponds to question two from the lesson plan task handout. Carrie was missing a representation of question 1 and 6a,

and Kathleen was only missing question 3a. What is important to note is that both participants have included a step for almost every question on the task handout. Together, their steps represent all phases of the worksheet, and most of the steps are listed with similar wording.

Table 27

Comparison of Carrie and Kathleen's Simulation Steps Connection to Lesson Plan

Carrie	Kathleen	Lesson Plan Handout for Task B
Not listed	1. Read the task and make a guess regarding thoughts on the outcome.	1. If 20 students took an exam, 10 on white paper and 10 on yellow paper, how do you think the average score from students who took the exam on yellow paper would compare to average score of the students who took the exam on white paper? Why?
1. Form various hypotheses about the scenario.	2. Determine the hypotheses that can be made.	2. How many different hypotheses could we make for this situation regarding the averages of scores of students who take the exam on yellow paper and on white paper? What are they?
2. Decide what the difference should be if the treatment has no effect.	Not listed	3a. In statistics, we typically subtract the average scores from two groups in order to compare them. If the color of the exam did not affect students' scores, what would be the most likely outcome (difference in the average scores) when this study is conducted with 20 participants?
3. Determine how much of a difference would be surprising if the treatment has no effect.	3. Make a statement regarding an outcome that would surprise. Create a boundary.	3b. Still assuming that the color of the exam did not affect students' scores (i.e. students would get the same score regardless of the color of the exam) what kind of results (difference in the average scores) would you not be surprised to see when this study is conducted with 20 participants?

Carrie	Kathleen	Lesson Plan Handout for Task B
4. Give the actual difference from the experiment.	4. Give the actual difference from the experiment.	4. For this experiment, the average test score for the yellow paper was _____ and the average test score for the white paper was _____. Therefore, the actual difference in the average scores of students who took the exam on yellow paper compared to students who took the exam on white paper was _____. If it is REALLY the case that the color of the exam doesn't matter, do you find the teacher's result surprising? Why or why not?
5. Perform a simulation just to generate some possible "chance" differences.	5. Design a simulation: roll dice, draw forma lot, random number generator, Table A? State precisely this design. 6. Carry out the simulation and record results.	5. Determine a plan for simulating the study.
	7. Discuss results (summarize)	6a. From your results, does it seem like the results obtained by the teacher would be surprising? Explain.
6. Use technology to simulate taking many samples and calculating the difference	8. Simulate with technology and compare results.	6b. Now, we will use technology to simulate this experiment many, many times under the assumption that the null hypothesis is true. Based on this simulation how surprising are the actual results of this study? Explain your reasoning
7. Calculate what percent of the time the difference obtained occurs.	9. State the *p*-value (part one of Kathleen's step 9).	6c. Based on the results of the simulation, how likely is a difference of 6.3 or greater? Explain.
8. Determine if this percent is high enough to be "surprising" or if it is due to chance.	9. write a conclusion (second part of step 9)	6d. Based on our simulations, what conclusion should the teacher draw? Justify your conclusion.

Chapter Summary

In this chapter I presented the result of a multiple-case study concerning how simulation tasks impacted teachers' understanding of hypothesis testing. Additionally, I shared how these teachers understood simulations and how they connected simulation and traditional approaches. The results showed that using simulations for hypothesis testing has a positive impact on how teachers understand hypothesis testing. The KDUs of the logic of hypothesis testing and probabilistic nature of hypothesis testing were the most influenced by the tasks. Carrie's knowledge was changed more than Kathleen's; however, Carrie's content knowledge was initially weaker. Carrie advanced her ability to articulate the reasoning behind a hypothesis test and deepen her understanding of *p*-value. Understanding when to reject or fail to reject based on the *p*-value is critical for hypothesis testing, and these tasks helped Carrie understand why a small *p*-value results in rejecting the null hypothesis. Also promising was Carrie's shift in focus from procedures to concepts. The simulation tasks also positively impacted a veteran teacher with a well-developed understanding of hypothesis testing. Kathleen deepened her understanding of the logic of hypothesis testing and strengthened her understanding of the significance level. Finally, the results from this study showed the direct impact of the lesson plan design on how the participants understood simulations in terms of the six-phases, which were designed to promote a deeper understanding of hypothesis testing. A summary and discussion of the results from this study will be shared in the next chapter.

CHAPTER FIVE: SUMMARY AND DISCUSSION

Introduction

There is a call for the educational community to produce statistically literate citizens (Franklin et al., 2015). However, teaching statistics has become the responsibility of mathematics teachers, who are generally recognized as being underprepared to teach this subject (Franklin, 2013). Also, research indicates that both teachers and students struggle with many of the difficult concepts found in an introductory statistics class (Harradine et al., 2011). Reform movements in statistics education have advocated for the use of technology and simulations to teach many of these abstract concepts (Franklin et al., 2015; Franklin et al., 2007). This dissertation sought to contribute to the knowledge base of using simulations in statistics education. In this chapter, I will provide a summary and discussion of this study. I will begin with an overview of the research problem and review of the methodology. Next, I will provide a summary of the results and conclude with a discussion of these results, along with recommendations of areas for future research.

Research Problem

One area of statistics that is particularly difficult is the topic of inference. At the high school level, inference topics, such as hypothesis testing, comprise a large portion of both AP and similar non-AP statistics curriculum. At the college level, simulations have gained much popularity for teaching inferential topics. Therefore, more knowledge regarding high school statistics teachers' understanding of hypothesis testing and the impact of using simulations on their knowledge would benefit the educational community in helping to plan training that these teachers need. Additionally, if high school statistics

teachers are going to use simulation approaches in their own classroom to foster a deeper understanding of hypothesis testing, it is important to know how they understand these techniques and how they connect traditional and simulation approaches.

The purpose of this study was to gain insight into how using simulations for hypothesis testing influences high school statistics teachers' understanding of hypothesis testing and how they understand simulations and connect that understanding to a traditional hypothesis test approach. My first research question was, "How does engaging in simulation tasks for hypothesis testing influence high school statistics teachers' understanding of traditional hypothesis testing?" To answer this question, my focus was on the teachers' CCK, which is their basic knowledge of hypothesis testing. However, I was also interested in assessing the content knowledge regarding simulations, and how those approaches are connected, that would allow teachers to use simulations to develop an understanding of hypothesis testing in their own classroom. This ability to use their knowledge of simulations and connecting approaches to promote a deeper understanding of hypothesis testing in their students corresponds to their SCK. Therefore, my second research question was, "How do simulation tasks influence high school statistics teachers' understanding of simulations and how do they make connections between traditional and simulation approaches for hypothesis testing?" The methodology that was used to answer these research questions will be reviewed next.

Review of Methodology

An explanatory multiple case study was used to answer the research questions of interest. The criteria for selection of participants was that they possessed at least some knowledge of hypothesis testing and that they had not engaged in simulation tasks for

hypothesis testing previously. Initially three participants were selected, and I collected data for all of them. However, results for one of the participants were not reported due to the discovery after data collection that he did not represent my population of interest. He had received training with simulations for hypothesis testing, but he did not realize this at the time because of different terminology used by myself and the trainer when describing this approach.

Pre-data was collected prior to the participants engaging in the simulation tasks. The participants took a pre-CAOS test and pre-open-ended response survey. Additionally, I interviewed each participant. The purpose of this pre-data was to provide information regarding their CCK for hypothesis testing prior to the simulation tasks. Next, I engaged the participants in three simulation tasks as a group approximately one week apart. I video recorded these tasks and had the participants complete a handout as they discussed each question and progressed through the lesson. Also, after each task, participants individually completed a post-task reflection. The purpose of these reflections was to gain insight into how the participants understood simulations and how they connected approaches. Finally, after completing all tasks, the participants took a post-CAOS test and post-open-ended survey, and I conducted a post-interview. These post-data pieces were designed to reassess the participants' CCK to determine changes in understanding.

I analyzed the data in stages, beginning with a deductive approach to classify data based on categories of CCK and SCK, along with an inductive approach used to explore the data and find factors which influenced changes in understandings. Next, I focused on using the analytical framework, which was designed to assess the participants'

understanding for each KDU category. The pre-data was organized and coded to classify the participant's CCK based on the five KDUs for hypothesis testing from this study's theoretical framework, which was created based on the literature review. Corresponding data pieces were compared with understandings listed in the analytical framework to determine if the participant possessed this understanding, and I developed narratives which summarized each participants' CCK organized by the KDUs. The post-data was analyzed and in the same manner, compared with the pre-data, and used to provide a descriptive narrative. This comparison determined the changes in CCK after engaging in the simulation tasks, which was the focus of research question one. Additionally, an inductive analysis using open-coding, analytical memos, and the production of the descriptive narratives was conducted to produce overall themes for research question one.

To answer research question two concerning the participants' SCK, I analyzed the post-task reflection pieces. The piece of the analytical framework that was used to assess the participants' understanding of simulations and how they connected approaches was comprised of Lane-Getaz and Zieffler's (2006) SPM model and their connecting approaches' model. I created tables to show a comparison of the participants' simulation steps and connection of approaches to their models. Additionally, I produced detailed narratives for all post-task reflection pieces to gain additional insights to answer research question two. Next, through deep reflection obtained from open-coding, analytical memos, construction of narratives, and rereading of all data pieces, I determined an overall theme for research question two. Finally, I used comparison tables of the pre- and post-data, changes in CCK, understanding of simulations, connection of approaches, and

themes to complete a cross-case analysis. A summary of these results will be provided next.

Summary of Results

In Chapter Four, I reported the descriptive narratives, which were produced from the multi-stage rigorous data analysis process. First, I presented the task narratives to allow the reader to gain insight into how the participants engaged in the simulation tasks. Next, to answer research question one concerning how the participants' content knowledge concerning hypothesis testing was influenced by the tasks, detailed narratives describing each case's CCK, organized by KDUs, were presented for both the pre- and post-data, highlighting these changes. To answer research question two, I presented the narratives concerning the post-task reflections, which detailed how the participants understood simulations and how they connected approaches. For each case, overall themes provided insights into the influencing factors for changes in the participants' CCK and how they understood simulations. Finally, a cross-case analysis showed the similarities and differences for the two cases. I will provide a summary of these results in this section.

Carrie's pre-data indicated that she viewed hypothesis testing in terms of steps and procedures. Additionally, unless she had access to these steps, she was unable to work traditional hypothesis test problems. Her pre-CAOS and pre-open-ended survey indicated that she had a basic grasp of many of the definitions related to hypothesis testing but lacked a robust understanding of such topics as p-value. Her pre-interview confirmed her lack of an in-depth understanding of many of these topics. For example, Carrie incorrectly described that a large *p*-value would result in rejecting the null

hypothesis because you were likely to obtain those results again. Also, she did not understand the relationship between the significance level and Type I error. In fact, she did not acknowledge that a mistake could be made and referenced accepting the null hypothesis.

However, the post-data revealed that many of these misconceptions were corrected. Her description of hypothesis testing had completely moved away from a procedural viewpoint. She showed a greater appreciation for the indirect reasoning behind a hypothesis test with her descriptions. This appreciation and understanding of the indirect reasoning of a hypothesis test was also shown when she logically thought through how to set up competing hypotheses when working a simulation problem in the post-interview. Also, her understanding of *p*-value, significance level, and Type I Error showed great changes. She could correctly explain why a small *p*-value resulted in rejecting the null hypothesis. She also acknowledged that mistakes could be made and knew that the Type I Error rate was determined by the significance level.

The simulation tasks allowed Carrie to visualize both the *p*-value and the significance level on the simulated sampling distribution. Carrie no longer had to perform complicated computations or rely on memorizing when to reject or fail to reject. In Carrie's own words, "The whole process became clear to me from writing the hypothesis to determining what the conclusion would be" (Post-open-ended question survey, November 27th, 2017), and "So, the simulation just gives us a quick and easy way to see the visualization of the distribution and to quickly find the probability of that value or less happening" (Post-interview, December 18th, 2017). The comparison of the pre- and post-data served to answer research question one by showing what aspects of Carrie's

CCK were changed after engaging in the simulation tasks. A further inductive analysis produced two overall themes for research question one. First, simulation tasks focus on concepts and logic instead of procedures, which develops the KDU of the logic of hypothesis testing. Second, simulations focus on visualizations, which help develop the KDU of probabilistic nature.

Research question two concerned understanding of simulations and connecting approaches. I had used Lane-Getaz and Zieffler's (2006) SPM to show an understanding of simulations as a three-tiered model. Additionally, I used Lane-Getaz and Zieffler's (2006) model for connecting approaches to assess Carrie's understanding of how the two approaches were related. Tables showing a comparison of Carrie's simulation steps and her connection of approaches to their models was provided in Chapter Four. These steps were apparent in Carrie's models that she constructed, however, she expanded these steps to consider such elements as thinking about the most likely outcome if the null hypothesis were true and what types of results would also not be surprising. This incorporation acknowledges the importance of variability. Other important considerations of expanding these simulation steps will be shared later in this chapter. As shown in Chapter Four, Carrie's expanded steps could be directly aligned with questions used in the simulation tasks. Therefore, for the second research question, the overall theme was that understanding of simulations and the connection of simulation and traditional approaches was influenced by the lesson plan design.

The second participant, Kathleen, showed a deeper understanding of hypothesis testing than Carrie did, as indicated by the pre-data. Kathleen could correctly work out both one and two-sample hypothesis test problems without consulting sources. Also,

unlike Carrie, Kathleen knew that a small *p*-value was when one would reject the null hypothesis. She also knew how important the randomness condition was and was adamant that even though you did not reject the null hypothesis that did not mean that it was true. However, Kathleen still showed some deficits in the logic of hypothesis testing KDU by initially incorrectly answering the Pepsi problem, and she did not possess a deep understanding of the significance level, which is a component of the probabilistic nature KDU.

However, after engaging in the simulation tasks, Kathleen showed a greater appreciation for the logic behind a hypothesis test. She could quickly interpret the results of the simulation from the Pepsi problem and was able to design her own simulation for the university satisfaction problem. The role of indirect logic and assessing unusualness from the probability obtained from the simulated sampling distribution was clear to Kathleen. Additionally, Kathleen gained a deeper understanding of the KDU probabilistic nature by being able to articulate the role of the significance level and how this was related to a Type I Error. The inductive analysis of the data indicated that it was the influence of the simulations' focus on concepts and logic and the visualization aspect of simulations that helped Kathleen develop her understanding of the logic of hypothesis testing and probabilistic nature KDUs, leading to the same overall themes for research question one as Carrie.

For research question two, I also provided tables comparing how Kathleen understood simulations and connected approaches to Lane-Getaz and Zieffler's (2006) models. Like Carrie, her models were expanded and seemed to be influenced by the lesson plan. This led to the same overall theme for research question two for Kathleen,

which was that understanding of simulations and the connection of simulation and traditional approaches was influenced by the lesson plan design.

The cross-case analysis revealed several similarities and difference. First, the initial content knowledge for each participant was different. Kathleen showed a deeper understanding of each of the KDUs for CCK as evidenced by the pre-data. Therefore, Carrie possessed more areas that she could show growth. According to the post-data, Carrie showed changes in her understanding in the category of indirect reasoning for the logic of hypothesis testing KDU and for categories of Type I error, significance level, and *p*-value for the probabilistic nature KDU. However, Kathleen only showed changes in her understanding of the indirect reasoning category for the logic of hypothesis testing KDU and of significance level for the probabilistic nature KDU. Thus, Carrie's additional categories for changes in understanding were Type I Error and *p*-value, but these were not areas that the pre-data indicated were lack of understandings for Kathleen. Therefore, although the types of understanding influenced were slightly different, the overall themes were the same for both participants for research question one. The data was even more consistent in terms of research question two. The simulation steps and how they connected approaches were very similar and the same overall theme for research question two was established.

Discussion of the Results

In this section I will discuss how the results of this study provided insights and contributions to the statistics educational community. First, I will share how the results are connected to the literature through the contribution to existing theory regarding teacher knowledge of hypothesis testing, understanding the logic of hypothesis testing,

and prerequisite knowledge needed to understand hypothesis testing. Next, I will discuss implications for the practice of using simulations to develop an understanding of hypothesis testing. Additionally, I will use the results to contribute to theory by expanding on models for simulations and for connecting approaches. Finally, I will offer suggestions for areas of future research.

Connections to the Literature

Teacher Knowledge of Hypothesis Testing. As reported in Chapter Two, teachers may be able to perform the computations of a hypothesis but lack an understanding of the reasoning behind these steps (Harradine et al., 2011). Additionally, teachers may hold some of the same misconceptions that students do regarding hypothesis testing (Harradine et al., 2011). My two participants had a combined teaching experience of 29 years, including a combined 14 years of teaching statistics. Yet, the pre-data showed that they still did not possess a robust understanding for all aspects of hypothesis testing, which aligns with the literature. This fact is not necessarily surprising, because the two teachers in this study felt that they had not been prepared to teach this subject. Both teachers indicated that they had not been taught many of the topics that they are expected to teach in their respective statistics classes. Therefore, this study provides further evidence for the lack of preparedness of our mathematics teachers to teach statistics. Specifically, the teachers in this study were lacking in the areas of the logic of hypothesis testing and probabilistic nature. Next, I will address how the participants' understanding of the logic of hypothesis testing aligned with the logic of hypothesis testing model created by Thompson et al. (2007) before and after engaging in the simulation tasks.

Commitment to the Null. Thompson et al. (2007) created a model for teachers' understanding of the logic of hypothesis testing (see Figure 2). This model showed four types of decisions that teachers from their study made. Two of the decisions follow the logic of hypothesis testing. These were that the null hypothesis is not rejected because the outcome is not unusual and that the null hypothesis is rejected because the outcome is unusual. However, two other types of thinking were evident that did not follow the logic of hypothesis testing. One was to not reject the null hypothesis because the results may have been biased. The other was to not reject the null hypothesis because if the sample has any chance of occurring under the null hypothesis model, then one should not reject the null hypothesis. These last two types of decision making, which Thompson et al. (2007) referred to as commitment to the null, were displayed in the pre-data by my participants.

In the pre- interview, I used the Pepsi problem found in the study by Thompson et al. (2007). This problem used a simulated sampling distribution which was created on the null hypothesis assumption that 50% of the population prefer Pepsi. The question then asks if you would conclude that more people prefer Pepsi if your sample data produced a statistic of 60%. On the sampling distribution that I showed my participants, 60% or higher occurred only 2.5% of the time. Initially, both participants concluded that there was not a majority that preferred Pepsi. Although, later in the interview Kathleen did switch her answer when relating the results to the traditional hypothesis step of rejecting the null when the *p*-value is smaller than alpha. Carrie said, "So, I mean that's what this one individual sample is and in that sample there was more people that favorite Pepsi, but no, I wouldn't say in the population" (Pre-interview, October 17th, 2017), and Kathleen

said, “So, it would be really hard to argue. It’s almost like this one was a fluke. Like it can happen” (Pre-interview October 19th, 2018). For both participants because the result could have happened, they did not believe that they had evidence for the alternative. This showed a commitment to the null hypothesis by both participants.

The other type of thinking explained by Thompson et al. (2007) that showed a commitment to the null hypothesis was also evident in Kathleen’s pre-open-ended response survey. This problem asked if she believed a claim that there was an 80% approval rating, if a sample of size 20 produced a 60% approval rating. Kathleen wrote, “Yes. If the survey is conducted by other students, there may be response bias. One survey is not enough (especially with 20 responses) to discredit the claim” (Pre-Open-Ended Response Survey, October 5th, 2017). Her statement is representative of the other commitment to the null decision in which teachers do not reject the null because they believe the sample size is too small or that there may be bias.

However, in the post-interview, neither participant showed evidence of the commitment to the null hypothesis type of thinking. For the Pepsi problem, both participants made the decision that the outcome was unusual and that more people preferred Pepsi. Also, for the approval rating problem in the post-open-ended response survey, Kathleen wrote, “No, because even though this sample is small, I believe it may be representative of the actual sentiments of the graduate population” (November 27th, 2007). As reported in Chapter Four, the simulation tasks focus on concepts and logic instead of procedures, which develops the KDU of the logic of hypothesis testing. After engaging in the tasks, participants in this study transitioned away from the commitment to the null types of decisions. One component that seemed to be impactful for this result

was having the participants discuss what types of results would surprise them before conducting the simulation. This early acknowledgement of variability may be a critical step to incorporate into simulations for hypothesis testing designed from a pedagogical perspective. As I will mention in the future research section, the connection of having participants state expected results and moving away from commitment to the null type of thinking is worth investigating.

Pre-requisite Knowledge. Another important aspect from the literature regarding hypothesis testing that was evident in this study was regarding pre-requisite knowledge. Garfield (2004) investigated the use of simulations to enhance students' understanding of sampling distributions in college. By the end of the study, some students still struggled with the concepts. Garfield found that these students often lacked the prerequisite vocabulary and understanding of distribution, variability, and models. Additionally, Saldanha and Thompson (2002) found that students who could interpret simulation results accurately had developed what they referred to as a multiplicative conception of sampling. Some affordances of this type of reasoning for hypothesis testing were the ability to distinguish between the three levels of population, sample, and sampling distribution and to refer to the simulation results in terms of a percent of sample statistics.

From the pre-data, the study's participants were able to correctly define most terms and showed at least a basic understanding that there was a difference among the levels of population, sample, and sampling distribution. However, although it seemed to be clear that all three participants knew that three distinct levels existed, neither participant referenced the sampling distribution in terms of the distribution of values taken by a statistic, which Saldanha and Thompson (2002) stated is done by students who

have developed multiplicative reasoning. In the pre-open-ended response survey (October 5th, 2017), Kathleen wrote that the sampling distribution was "a group of all possible samples taken from a population," and Carrie defined it as "all possible samples of a certain size from the population". Also, during the second task, when I asked Carrie what the simulated sampling distribution was, she said, "I mean this is just a simulation. This is just one thousand six hundred and eighty trials of this experiment. When we do the calculations the old way, that's not just simulations that the sampling distribution, that's all possible samples" (Task B, November 15th, 2018). These same types of definitions focusing on all samples versus all statistics calculated from the samples were expressed in the post-data as well. However, based on the participants being able to correctly interpret the results of the simulations, I believe that they must have possessed a multiplicative conception of sampling, but they were just not careful or expert enough in the description of a sampling distribution. Evidence for this was apparent during the simulation tasks. When performing the simulations, they recorded the statistic obtained from their sample. Additionally, their conclusions were stated in terms of theme the statistic on the simulated sampling distribution and determining the probability of obtaining a statistic that extreme or more extreme. Additive versus multiplicative reasoning would mean that they viewed a sampling distribution as a collection of individuals not statistics. Therefore, although the participants' data did not show an explicit definition of sampling distribution as a collection of statistics, I believe that they were using multiplicative reasoning by their ability to draw an appropriate conclusion based on the sampling distribution in terms of the statistic. Therefore, this study does show evidence that possessing statistical vocabulary and at least some form of

multiplicative sampling reasoning are important elements for successfully interpreting the results of a simulation. Additionally, it may be difficult to determine if someone possesses multiplicative reasoning just based on their definitions of sampling distribution.

Implications for Practice

Although the results from this study are promising for the use of simulation tasks to help develop an understanding of hypothesis testing, it should be noted that this is not a panacea for correcting all misconceptions and many factors influence the success of using these tasks. First, it is important to note that even though the participants did display some areas in which they lacked knowledge, the number of years teaching experience with statistics probably influenced how easily they were able to grasp many elements of the simulation tasks. Through my personal experience of using simulation tasks in the classroom and from the literature showing that students engaging in simulation tasks are not always able to interpret results from these tasks (e.g. Saldanha & Thompson, 2002), it is not reasonable to assume that students and even teachers will readily be able to interpret the results of a simulation and be able to make connections to a traditional approach. As noted in the previous section, some prerequisite knowledge concerning vocabulary and at least a beginning development of multiplicative sampling reasoning is probably required. Additionally, how the simulation task unfolds and how the simulations are connected to the traditional approach are critical elements in terms of the type of knowledge that can be gained.

However, what was promising was that for both of the participants, even Kathleen who seemed to possess a strong understanding of the KDUs for hypothesis testing, gains in understanding were obtained, as evidenced by the post-data. The simulation tasks

decrease the amount of details, such as formulas and procedures, so that they could focus on the key concepts of hypothesis testing instead. Significance levels and *p*-values were visualized on the simulated sampling distribution making the probabilistic nature KDU easier to develop. Also, the indirect reasoning of hypothesis testing was able to take precedence over following the steps of a hypothesis test, making the logic of a hypothesis test become apparent. Based on the results from this study, I will now share what I believe are important contributions to models for simulations and for connecting approaches and how these models can be used.

Expansion of Simulation Steps

The lesson plan used for the simulation tasks for this study extended the simulation steps beyond the traditional three-tiered approach used to develop an understanding of traditional hypothesis testing found in the literature. As a result, the participants viewed using simulations for hypothesis testing as encompassing additional steps. Also, these steps seemed to help the participants make sense of traditional hypothesis testing. Therefore, I have created a Hypothesis Test Simulation Model (HTSM) with an expanded list of simulation steps, from a pedagogical perspective of using simulations to develop a deeper understanding of a traditional hypothesis testing, as shown below (see Figure 15).

Hypothesis Test Simulation Model (HTSM)

1. Establish population parameters in context.

2. List possible competing models.

3. Identify the null model in terms of a specific parameter value (most likely outcome).

4. Consider spread of the null model (what do you expect due to chance).

5. Gather evidence from random sample or randomized experiment.

6. Design simulation in terms of the repeatable action and perform tactile simulation.

7. Build simulated sampling distribution with technology.

8. Determine estimated *p*-value to assess unusualness and make a conclusion.

Figure 15. Hypothesis Test Simulation Model (HTSM)

However, to use the HTSM to help develop an understanding of traditional hypothesis testing, special attention should be given to how to use these steps in tasks. Therefore, I have created a table which shows how the steps from the HTSM are connected to possible lesson plan phases, along with notes and affordances for the steps (see Table 28). This table is designed to assist anyone using simulation tasks to develop a

deeper understanding of hypothesis testing. It could be used in either a professional development setting or by a teacher in the classroom.

Table 28

Alignment of Simulation Steps, Lesson Plan, Notes, and Affordances

HTSM	Lesson Plan Phases	Notes	Affordance for Learning Outcomes
1. Establish population parameters in context.	Commitment to a position in a rich context.	Discussion should focus on identifying the parameter of interest.	Ability to make sense of the problem and recognize parameter of interest, focused on context, is developed.
2. List possible competing models.	Statement of possible hypotheses.	List hypotheses in words, stated in terms of the parameter of interest. Hypotheses should be competing.	Indirect reasoning of the KDU logic of hypotheses testing is developed.
3. Identify the null model in terms of a specific parameter value (most likely outcome).	Statement of most likely result assuming the null hypothesis is true.	Focus should be on using the model that competes with what you are trying to gather evidence for. Most likely outcome stated in terms of the parameter of interest.	Mean of sampling distribution is determined. Recognition of the population level from three-tiers of sampling. Developing understanding of indirect reasoning.
4. Consider spread of the null model (what do you expect due to chance).	Statement of expected results assuming the null hypothesis is true.	Begin to think about what would or would not be enough evidence to reject the null hypothesis.	Understanding of the variability of the statistic is developed.
5. Gather Evidence from random sample or	Revelation of study results.	Focus on this being from one sample or one experiment and what this means in	Acknowledgement that you are using sample data to make inferences

HTSM		Lesson Plan Phases	Notes	Affordance for Learning Outcomes
	randomized experiment.		terms of the type of conclusion you can draw.	about a population or to determine causation.
6.	Design simulation in terms of the repeatable action and perform tactile simulation.	Simulation.	Focus should be on mimicking the random assignment or sampling and calculating the statistic.	Recognition of what is repeatable and the statistic of interest. Recognition of the sample level of the three-tiers of sampling. Development of multiplicative reasoning.
7.	Build simulated sampling distribution with technology.	Simulation.	Building of the distribution of the statistic by repeating the random assignment or sampling should be emphasized.	Recognition of the level of sampling distribution from the three-tiers to develop multiplicative conception of sampling.
8.	Determine estimated p-value to assess unusualness and make a conclusion.	Make a conclusion.	Be explicit about the role of the sampling distribution as a probability model. Convincing evidence if your sample statistic was surprising enough. Highlight significance level.	Probabilistic nature of hypothesis testing emphasized. Significance level and Type I Error can be identified.

From the lesson plan design (See Strayer & Matuszewski, 2016) used to create the tasks, the six phases are (1) commitment to a position in a rich context, (2) statement of possible hypotheses, (3) statement of expected results assuming the null hypothesis is true, (4) revelation of study results, (5) simulation under the null hypothesis, (6) making a conclusion. However, I have separated phase three into two parts by first stating the most likely result and then listing the other types of expected results. I did this because the task

handout separated this phase into two questions and each part addresses different elements in terms of the simulation, as noted in the table. Additionally, phase five of conducting the simulation was divided into the tactile simulation and simulation using technology. Again, this aligns with the task handout in which both types of simulations were conducted, and each part focuses on different elements. For each simulation step and corresponding lesson plan phase, I provided notes regarding specific details that should be focused on during the task, along with the affordance for the learning outcome of hypothesis testing that can be gained listed in the last column. Each step provides opportunities to develop a more robust understanding of hypothesis testing, especially when attending to specific elements. I will describe each step in detail next.

Step one of the HTSM is to establish population parameters in context. In the lesson plan this is done by having the learner commit to a position in a rich context. For example, in the first task, participants were asked how many of the children they believed chose the helper toy out of 20. Therefore, the parameter of interest was the number of children who would choose the helper toy if the experiment were repeated with 20 children. The question could also be worded so that the parameter of interest was a population proportion. To determine the parameter of interest, learners must make sense of the problem by focusing on the context, which is critical for statistics. Kathleen commented that discussing the problem was what was missing in traditional hypothesis test problems in textbooks.

The second step of the HTSM is to list possible competing models, which corresponds to statement of possible hypotheses on the lesson plan. During this stage, an awareness of the indirect reasoning of hypothesis testing can begin to develop by

focusing on determining competing hypotheses. The teacher can then identify which one is the null hypothesis, emphasizing this should be in competition to what you are trying to gather evidence of with your study. For Carrie, this step allowed her to stop focusing on procedures and begin to consider the overall concept. This step corresponded to question two on the handout, and as reported in Chapter Four, Carrie showed a transition from the first task for this question of using traditional hypothesis test notation to the last task where she wrote out the hypotheses based solely on the context.

After listing the hypotheses and identifying the null hypothesis, the third simulation step clarifies the population model that will be used by identifying the null model in terms of a specific parameter value. In the lesson plan this is accomplished by asking what the most likely outcome would be if the null hypothesis is true. To prevent commitment to the null type of thinking, it is important to emphasize that this is not what you are trying to gather evidence to establish. Additionally, this specific parameter value should be the mean of the simulated sampling distribution, which also allows one to focus on the population level from the three tiers of sampling described by Saldanha and Thompson (2002). Before engaging in the tasks, both participants showed an inclination to commit to the null hypothesis. Therefore, it is critical that for this step the null hypothesis is seen as an assumption model, not what is true. After the tasks, both participants viewed the null model as an assumption and a way to determine how convincing the evidence was.

The next step is to consider the spread of the null model, by asking the learners what types of results would not surprise them. This allows the learner to acknowledge the variability that should be expected due to chance. Additionally, this begins a discussion

of how much evidence would be required to convince you to reject the null hypothesis. This additional step seemed to help the participants move away from the commitment to the null type of thinking. However, as teachers, my participants possessed the prerequisite knowledge of expecting their results to vary. For students, this prerequisite knowledge may need to be established or the simulation may help them develop this understanding.

In step five of the HTSM, evidence is gathered from a random sample or experiment. In the lesson plan this is achieved by revealing the study's results. Other simulation models did not mention the importance of checking the random condition, which still should be done with a simulation approach. At this stage, the focus should be on the fact that you are using the results of this one sample or experiment to infer something about a population of interest or to infer causation. This can only be achieved if data was obtained from a random sample or randomized experiment. The learner can then achieve a greater appreciation for a hypothesis test and what type of inference can be made. This stage can also serve to continue interest, because even the teachers in this study were very excited to see how close their predictions were from the first question of the lesson plan.

The sixth step of the HTSM, which is to design the simulation in terms of the repeatable action, corresponds to tactile simulation on the lesson plan. By having the learner design the actual simulation he or she must explicitly state what is being repeated and what statistics that he or she will calculate. This purposefully focuses on the sample level from the three-tiers of sampling. Additionally, by having the learner design the simulation, this may help he or she to develop their multiplicative reasoning by

emphasizing that one is collecting a group of statistics, not individuals, and allow the visualization of the distribution to be more impactful. For my participants, this step took little prompting from me for the first two tasks. The design of the simulation for the third task took a little more time, but they were still successful. Again, as teachers, this step was probably more easily accomplished by having certain prerequisite knowledge. Both participants had taught how to design simulations to estimate probabilities. However, special care and scaffolding will be needed for students to ensure they understand what is being repeated and what the statistic of interest is so that connections can be made when using technology.

The seventh step is building a simulated sampling distribution. On the lesson plan, this corresponds to using technology to perform the simulation. The learner should witness the technology performing several trials and relate that to what he or she just did with the tactile simulation. This provides an emphasis on the sampling distribution as a collection of statistics. Also, the sampling distribution is seen as its own tier from the three-tiers of sampling and can be used to develop a multiplicative conception of sampling. This step was interesting for my participants. Kathleen said, "And when you can watch this build that's just incredible. Yeah, I love that" (Post-interview, December 19th, 2017). Kathleen mentioned several times how amazing it was to watch the sampling distribution build. For Carrie, she thought the simulations were more real by being able to see the sampling distribution. She wrote about simulations on her post-task reflection after Task B, "We just have a 'real' normal curve rather than the one in our books" (November 15th, 2017). By this statement, it is not clear if Carrie thought the simulation was more real because it was constructed using the real data from the yellow-white tasks

or if the theoretical sampling distribution never made sense to her so was never real in her mind. However, for both participants this building of the sampling distribution offered an opportunity to help their students make sense of the difficult concept of sampling distribution.

The last step of the HTSM is to determine the estimated *p*-value to assess unusualness and make a conclusion, which is simply make a conclusion from the lesson plan phase. This step offers a unique opportunity to develop the probabilistic nature KDU. The role of the simulated sampling distribution as a probability model should be made explicit. To do this, the sample statistic should be located on the distribution and the probability of obtaining a value that extreme or more extreme should be estimated from the distribution. Also, as Kathleen noted, this allows one to visualize why one should reject when the *p*-value is small, because one can see how few data points are in the tails. Therefore, statistics that fall in the tail are surprising and are evidence that the alternative hypothesis is true. The probabilistic nature KDU can be further developed in this step by using technology to highlight the area that corresponds to the significance level. In this study, this allowed the participants to visualize the rejection region and to see how the probability of committing a Type I Error is determined. Also, by considering the context, one can develop the idea of the importance of determining if the significance level should be larger or smaller. As seen in the yellow-white task, if one were a student, then committing a Type I Error, which would mean determining that yellow paper helps your test score when it does not, is not major consequence. However, if one is a teacher having to spend extra money on colored paper, one would be less willing to make this type of mistake.

As just described, by expanding the simulation steps and focusing on critical elements during the simulation tasks, many affordances for student learning are achieved. However, I believe more gains in understanding may be achieved by making the connection between the simulation and traditional approaches more explicit. Also, by having a model for linking these approaches, designers of professional development for teachers and classroom teachers can have access to a resource to help develop an understanding of hypothesis testing using simulations. I will describe my model for connecting approaches next.

Connection of Approaches

A need for a model connecting approaches was shown by the difficulty my participants had with this task. First, it was not until the third post-task reflection where I provided them with a table to list the steps side by side that my participants explicitly connected the two approaches, even though they were asked to do so after each of the first two tasks. Additionally, the participants in this study did not list all the necessary steps to conduct a traditional hypothesis test problem. Specifically, Carrie omitted "checking conditions" for a traditional approach, and Kathleen did not list "calculate a test statistic." Neither participant mentioned the role of a theoretical sampling distribution. Also, although Kathleen listed checking conditions, she was unsure how this related to the simulation approach. She wrote on her post-task C reflection for the simulation step that was connected to the traditional hypothesis step of check conditions, "Typically not as much time spent here. This still needs to be done, but through understanding of the problem, this can happen" (November 27th, 2017). However, the simulated sampling distribution does not require for the sampling distribution to be

normal to obtain a *p*-value and is unnecessary to check. Also, although Lane-Getaz and Zieffler (2006) provided a model connecting approaches, their simulations steps do not encompass my expanded list of simulation steps that I just described. Therefore, the following table (see Table 29) aligns each of my simulation pedagogical steps to a traditional approach.

Table 29

Model for Connecting Simulation and Traditional Hypothesis Test Approaches

Simulation Steps from HTSM	Traditional Hypothesis Test Steps
Establish population parameter in context.	Define parameter of interest
List possible competing models.	Statement of null and alternative hypothesis.
Identify the null model in terms of a specific parameter value (most likely outcome).	Mean of theoretical sampling distribution used in calculating *z* or *t* test statistic
Consider spread of the null model (what do you expect due to chance).	Standard Error used in calculating *z* or *t* test statistic.
Gather evidence from random sample or random experiment	Determine sample statistic and calculate *z* or *t* test statistic.
Design simulation in terms of the repeatable action and perform tactile simulation.	N/A
Build simulated sampling distribution with technology.	Check conditions to determine validity of using theoretical sampling distribution.
Determine estimated *p*-value to assess unusualness and make a conclusion.	Determine *p*-value, compare to alpha, and make a conclusion. (or compare test statistic)

Many textbooks list a traditional approach with only four steps. For example, Starnes et al. (2010) listed the following steps.

1. State: What hypothesis do you want to test, and at what significance level. Define any parameters you use.
2. Plan: Choose the appropriate method and check conditions.
3. Do: Perform calculations. Compute the test statistic and find the *p*-value.
4. Conclude: Make a decision about the hypothesis in the context of the problem.

Notice that each step is represented in my table, but I also listed determine the mean and standard error for the theoretical sampling distribution separate from calculating the test statistic. Students often memorize the formula for a test statistic but do not understand what it means (Harradine et al., 2011). By making the pieces of the formula more explicit and connecting it to something they can see on the simulated sampling distribution may help learners understand the formula. Another important consideration is that the simulation step of considering the spread of the model by stating the types of results that would be expected is connected to the standard error of the traditional approach. The standard error is often not discussed much in the traditional approach and may even be hidden if using technology to calculate the test statistic. However, in the simulation approach, this step led to much discussion by the participants during the tasks. As I mentioned under the connections to literature section of this chapter, this element of discussing the types of results that would surprise the participants may have helped them move away from the commitment to the null type of thinking.

Also notable from this connecting approaches model is that designing the simulation in terms of the repeatable action and performing the tactile simulation is not

aligned with a traditional step. Because the simulation steps are from a pedagogical perspective, the inclusion of this step and its impact on learning is important to consider. The lesson plan design included this step to have the learner make sense of what the technology was doing, but there may be even more of an impact from a visualization perspective as will be discussed in the future research section.

Another important distinction when considering this model for connecting approaches is that the simulation steps are listed from a pedagogical perspective of helping to develop an understanding of traditional hypothesis testing. A simulation approach, purely from a statistical perspective, could be used to work any problem that can be solved using traditional methods. The power of using simulations to construct the sampling distribution is that you do not need to check conditions to determine the shape of your sampling distribution. For a traditional approach, conditions must be checked to determine if your theoretical model, based on a function, will be accurate. However, if conditions are met and the theoretical model can be determined, a traditional approach can be faster and more efficient to solve the problem of interest. The main idea is that if one is considering comparing a simulation and traditional approach to solve a problem statistically, the only difference is in the type of model used. The sampling distribution is either constructed using simulations or determined by a function. This idea was not apparent for my participants. In fact, Carrie mentioned that the simulated distribution was the "real" distribution. Neither the simulated or the theoretical sampling distribution are necessarily the real one. They are both different ways to model the distribution of the sample statistic from repeated sampling. Neither participant acknowledged the role of the theoretical sampling distribution in the traditional approach and did not seem to

understand that the *z*- or *t*-distribution is a model based on a function. Therefore, they were unable to connect this to something in the simulation approach. An understanding of these two types of models, which depend on the approach used, should be developed in teachers to help them gain a deeper understanding of hypothesis testing and to more effectively use simulations to foster this understanding in their own students.

In conclusion, using a simulation approach from a pedagogical perspective allows the logic of hypothesis testing to take precedence over formulas and tedious calculations. Determining the *p*-value and using it to make a decision becomes easier by seeing this value on the simulated sampling distribution. Simulations also naturally show the three levels of sampling and can help learners develop a multiplicative conception of sampling. However, if the goal is to understand traditional hypothesis testing, then more care and attention must be given to each simulation step and how it is connected to the traditional approach. Next, I will discuss how to build on the results from this study with a discussion of possible areas for future research.

Future Research

This study looked at how teachers' content knowledge was influenced after engaging in simulation tasks, how these tasks influenced how they understood simulations, and how they connected approaches. However, I did not look at how the changes in the teachers' understanding may have persisted over time or how engaging in these tasks may have affected their classroom practices. Therefore, future research might focus on the retention of knowledge gained because of these tasks and how teachers' classroom practices may be influenced. Specifically, do the teachers focus on discussions designed to highlight the logic of hypothesis testing, and can they critique different

simulation designs? Also, how do they develop an understanding of the different levels of sample and how do they encourage an understanding of the probabilistic KDU?

Also, in terms of the technology component, I did have the participants use the technology in the post-interview to answer a question using a simulation. Both participants were able to establish the parameters, run the simulation, and produce a sampling distribution that they used to make a conclusion based on the simulated p-value. However, I did not investigate the participants' use of technology in their own classroom. It may be beneficial to investigate the technological pedagogical content knowledge (TPACK), which is necessary for effectively using simulation technology for tasks designed to promote an understanding of hypothesis testing.

Additionally, from the results of this study, I proposed using an expanded list of simulation steps as given by the HTSM and incorporating an explicit model for connecting approaches when using simulation tasks to promote a deeper understanding of hypothesis testing. Therefore, research investigating the efficacy of these models in promoting understanding would be beneficial. Specifically, does the element of having participants acknowledge variability early in terms of stating expected results help transition away from the commitment to the null type of thinking? Additionally, what aspects of the tasks help participants make more sense of the visualization component? Traditional textbooks show a picture of the theoretical sampling distribution with the *p*-value shaded. However, neither participant indicated that this element had impacted their knowledge previously. By looking at the difference between the HTSM and traditional approach, a notable difference is that designing the simulation in terms of the repeatable action and performing the tactile simulation are not aligned with a step in the traditional

approach. Some researchers have suggested that the tactile element helps students make sense of what the technology is doing (e.g. Holcomb et al., 2010). However, no research has been conducted investigating the impact of designing the simulation in terms of the repeatable action. This element could have helped the participants understand and interpret the simulated sampling distribution, allowing the visualization component to be more impactful.

The HTSM and Connecting Approaches Model should also be investigated using different populations of interest, such as pre-service teachers, in-service teachers, and students would highlight the differences that may arise. Students would probably struggle at different stages of the tasks and would need different scaffolding techniques. Teachers may be ready for more in-depth investigations and not need as much time spent at different stages of the tasks.

Finally, although the focus of this study was on using a simulation approach as a pedagogical tool for developing an understanding of traditional hypothesis testing, a special note should be made regarding the difference in the model used for each approach. For a simulation approach, the model used to estimate the *p*-value was created by simulation. However, in a traditional approach, the model used is the theoretical sampling distribution, which is derived from a function. From analyzing the data, neither participant seemed to acknowledge or understand the role of the theoretical distribution in the traditional approach. Future research should investigate how to develop teachers' understanding of the theoretical sampling distribution as a model of the repeated sampling, which is derived from a function, and how this connects to creating a model through simulation.

Chapter Summary

Statistics is such a unique field, because it is not just about numbers or formulas. It is about predictions, patterns, and sense. One must use the context, such as dolphins, Pepsi, helper toys, and university approval rates, to make meaning from the data. In a classroom where the focus is on steps and procedures, these ideas are lost, and students are left thinking that statistics is a field that relies on a set of abstract computations that only few can understand. However, when the focus becomes on the big picture of hypothesis testing using simulations, a deeper understanding and appreciation for hypothesis testing can be achieved.

This study sought to contribute to the knowledge base of using simulations to help develop a deeper understanding of hypothesis testing for high school statistic teachers. Additionally, I investigated how simulation tasks influenced how these teachers understood simulations and connected traditional and simulation approaches. The results showed that both participants displayed a deeper understanding of traditional hypothesis testing after engaging in the simulation tasks. Carrie moved away from using procedures and began to focus on the logic of hypothesis testing. As a result, she showed an increased appreciation for the indirect reasoning behind a hypothesis test. Additionally, her understanding of *p*-value, Type I Error, and the significance level were all enhanced. Kathleen was able to develop her understanding of both the indirect reasoning of a hypothesis test and the significance level. The data indicated that the focus of simulations on concepts and the visualization aspect of these tasks helped the participants develop these understandings.

Also, by looking at how the participants understood simulations and connected approaches, this study contributed to the knowledge base of simulations for inference by offering an expanded model for simulations and a model for connecting traditional and simulation approaches. These models can be used to help design simulation tasks, which foster an understanding of essential concepts of hypothesis testing for not only students, but teachers as well. Additionally, this research prompted insights into how acknowledging variability early may help participants move away from commitment to the null type of thinking and how designing the simulation in terms of the repeatable action may enhance the visualization component. Finally, investigating how teachers understand the role of function in the theoretical sampling distribution and how this connects to the type of model used in the simulation approach are all fruitful areas for future research.

References

Antonioli, C., & Reveley, M. A. (2005). Randomised controlled trial of animal facilitated therapy with dolphins in the treatment of depression. *British Medical Journal*, *331*, 1231.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*, 389–407.

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning*, *2*, 75-97.

Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3-15). Dordrecht, The Netherlands: Kluwer Academic.

Bloomberg, L. D., & Volpe, M. (2012). *Completing your qualitative dissertation: A road map from beginning to end*. CA, US: Sage Publications.

Cobb, G. W. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education*, *1(1*),1-16.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, *104*(9), 801-823.

Collegeboard. (2010). AP Statistics course home page. Retrieved from https://apcentral.collegeboard.org/courses/ap-statistics/exam.

delMas, R. C. (2004). A comparison of mathematical and statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 79-95). Netherlands: Springer.

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

Erickson, T. (2006). Using simulation to learn about inference. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Bahia, Brazil.

Evans, J. S. B. (1989). *Bias in human reasoning: Causes and consequences*. Mahwah NJ: Lawrence Erlbaum Associates, Inc.

Franklin, C. (2013). Guest Editorial . . . Common Core State Standards and the Future of Teacher Preparation in Statistics. *The Mathematics Educator, 22*(2), 3-10.

Franklin, C. A., Bargagliotti, A. E., Case, C. A., Kader, G. D., Scheaffer, R. L. & Spangler, D. A. (2015) *The Statistical Education of Teachers*: Alexandria, VA: American Statistical Association.

Franklin, C., Hartlaub, B., Peck, R., Scheaffer, R., Thiel, D., & Tranbarger Freier, K. (2011). AP Statistics: Building bridges between high school and college statistics education. *The American Statistician, 65*, 177-182.

Franklin, C., & Kader, G. (July, 2010). *Models of teacher preparation designed around the GAISE framework.* Paper presented at the Eighth International Conference on Teaching Statistics, Ljubljana, Slovenia.

Franklin, C., Kader, G., Mewborn, D. S., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines and assessment for instruction in statistics education (GAISE) report: A pre-K-12 curriculum framework.* Alexandria, VA: American Statistical Association.

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education*, *19*, 44-63.

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review*, *75*, 372-396.

Garfield, J., delMas, R., & Chance, B. (2007). Using students' informal notions of variability to develop an understanding of formal measures of variability. In M.C. Lovett & *P*. Shah (Eds.), *Thinking with data*, (pp. 117-147). New York, NY: Lawrence Erlbaum Associates.

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

Gürbüz, R., & Birgin, O. (2012). The effect of computer-assisted teaching on remedying misconceptions: The case of the subject "probability". *Computers & Education*, *58*, 931-941.

Gonzalez, O. (2012). A framework…teachers. In D. Ben-Zvi and K. Maker (Eds.), *Teaching and learning statistics. Proceedings of Topic Study Group 12, 12th Internatinal Congress on Mathematical Education (ICME-12), Seoul, Korea.*

Groth, R. E. (2007). Toward a conceptualization of statistical knowledge for teaching. *Journal for Research in Mathematics Education*, 427-437.

Harradine, A., Batanero, C., & Rossman, A. (2011). Students and teachers' knowledge of sampling and inference. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics-challenges for teaching and teacher education* (pp. 235-246). The Netherlands: Springer.

Hirsch, L. S., & O'Donnell, A. M. (2001). Representativeness in statistical reasoning: Identifying and assessing misconceptions. *Journal of Statistics Education, 9,* 61-82.

Holcomb, J., Chance, B., Rossman, A., Tietjen, E., & Cobb, G. (2010). Introducing concepts of statistical inference via randomization tests. *Proceedings Data and context in statistics education: Towards an evidence-based society (ICOTS8), Voorburg, The Netherlands*.

Krauss, S., & Wassner, C. (2002). How significance tests should be presented to avoid the typical misinterpretations. In *Proceedings of the Sixth International Conference on Teaching Statistics.* Cape Town, South Africa: International Association for Statistics Education.

Lane, D. M., & Tang, *Z.* (2000). Effectiveness of simulation training on transfer of statistical concepts. *Journal of Educational Computing Research, 22,* 383-396.

Lane-Getaz, S. J. (2010). Linking the randomization test to reasoning about *p*-values and statistical significance. 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.

Lane-Getaz, S. J., & Zieffler, A. S. (2006). Using simulation to introduce inference: An active learning approach. *2006 Proceedings of the American Statistical Association*, Alexandria, VA: American Statistical Association.

Lee, H. S., Doerr, H. M., Tran, D., & Lovett, J. N. (2016). The role of probability in developing learners' models of simulation approaches to inference. *Statistics Education Research Journal*, *15*, 216-238.

Liu, Y., & Thompson, *P*. W. (2009). Mathematics teachers' understandings of proto-hypothesis testing. *Pedagogies: An International Journal*, *4*(2), 126-138.

Lovett, J. N., & Lee, H. S. (2017). New Standards Require Teaching More Statistics: Are Preservice Secondary Mathematics Teachers Ready?. *Journal of Teacher Education*, *68*(3), 299-311.

Makar, K., & Confrey, J. (2004). Secondary teachers' statistical reasoning in comparing two groups. In D. Ben-Zvi & J. Garfield (Eds.), *The challenges of developing statistical literacy, reasoning, and thinking* (pp. 353–374). Dordrecht, The Netherlands: Kluwer Academic.

Miles, M. B., Huberman, A. M., & Saldana, J. (2014). *Qualitative data analysis: A method sourcebook.* CA, US: Sage Publications.

Moritz, J. (2004). Reasoning about covariation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenges of developing statistical literacy, reasoning, and thinking* (pp. 227-256). Dordrecht, The Netherlands: Kluwer Academic.

National Council of Teachers of Mathematics. (2014). *Principles to actions: Ensuring mathematical success for all*. Reston, VA: Author.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. J. Kilpatrick, J. Swafford, and B. Findell (Eds.). Mathematics Learning Study Committee, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 175.

Noll, J. A. (2011). Graduate teaching assistants' statistical content knowledge of sampling. *Statistics Education Research Journal*, *10*(2), 48-74.

Patton, M. Q. (2015). *Qualitative evaluation and research methods*. Thousand Oaks, CA: Sage Publications.

Peck, R., Gould, R., & Miller, S. J. (2013). *Developing essential understanding of statistics for teaching mathematics in grades 9-12*. Reston, VA: National Council of Teachers of Mathematics.

Peters, S. A. (2009). *Developing an understanding of variation: AP Statistics teachers' perceptions and recollections of critical moments.* (Doctoral dissertation, The Pennsylvania State University).

Polya, G. (1945). *How to solve it: A new aspect of mathematical method*. Princeton, NJ: Princeton university press.

Reading, C., & Shaughnessy, J. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenges of developing statistical literacy, reasoning, and thinking* (pp. 201-226). Dordrecht, The Netherlands: Kluwer Academic.

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

Saldaña, J. (2016). *The coding manual for qualitative researchers.* Thousand Oaks, CA: Sage Publications.

Saldanha, L. A., McAllister, M. (2014, July). Using resampling and sampling variability in an applied context as a basis for making statistical inferences with confidence. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in Statistics Education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9),* Flagstaff, Arizona.

Saldanha, L. A., & Thompson, *P*. W. (2014). Conceptual issues in understanding the inner logic of statistical inference: Insights from two teaching experiments. *The Journal of Mathematical Behavior, 35,* 1-30.

Saldanha, L., & Thompson, *P*. (2002, October). Students' scheme-based conceptions of sampling and its relationship to statistical inference. *Proceedings of the Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education,* Athens, GA.

Scheaffer, R. L. (2006). Statistics and mathematics: On making a happy marriage. In G. F. Burrill & P. C. Elliot (Eds.), *Thinking and reasoning with data and chance*, (pp. 309-322). Reston, VA: National Council of Teachers of Mathematics.

Shaughnessy, J. M. (1977). Misconceptions of probability: An experiment with a small-group, activity-based, model building approach to introductory probability at the college level. *Educational Studies in Mathematics, 8,* 295-316.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester

Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 957-1010). Charlotte, NC: Information Age.

Smith, *T*. M. (2008). *An investigation into student understanding of statistical hypothesis testing*. University of Maryland, College Park.

Sotos, A. E. C., Vanhoof, S., Van den Noortgate, W., & Onghena, *P*. (2009). How confident are students in their misconceptions about hypothesis tests?. *Journal of Statistics Education*, *17*(2), 1-22.

Starnes, D. S., Yates, D., & Moore, D. S. (2010). *The practice of statistics*. New York, NY: Macmillan.

Statcrunch (n.d.). Randomization test for two means. Retrieved from https://www.statcrunch.com/5.0/example.php?example_id=99

Strayer, J., & Matuszewski, A. (2016). Statistical Literacy: Simulations with Dolphins. *Mathematics Teacher*, *109*, 606-611.

Thompson, *P*. W., Liu, Y., & Saldanha, L. A. (2007). Intricacies of statistical inference and teachers' understandings of them. *Thinking with Data*, 207-231.

Tintle, N. L., Rogers, A., Chance, B., Cobb, G., Rossman, A., Roy, S., & VanderStoep, J. (2014). Quantitative evidence for the use of simulation and randomization in the introductory statistics course. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in Statistics Education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9),* Flagstaff, Arizona.

Tintle, N., VanderStoep J., Holmes V-L., Quisenberry B., & Swanson *T*. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education*, 19(1), 1-25.

Watson, J. M. (2004). Developing reasoning about samples. In D.Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 277-294). Netherlands: Springer.

Watson, J. M., & Moritz, J. B. (2000). Developing concepts of sampling. *Journal for research in Mathematics Education, 31*, 44-70.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223-248.

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

APPENDICES

APPENDIX A: BACKGROUND SURVEY

1. How many years have you been teaching?
2. What degree(s) do you hold?
3. How many years have you taught statistics?
4. What type of statistics courses have you taught?
5. How many classes in statistics in college did you take?
6. How would you describe your teaching philosophy in a couple of sentences?
7. What other classes do you teach currently?
8. What other classes have you taught in the past?
9. What kind of training have you received to teach statistics?
10. How did you prepare on your own to teach this class?
11. Have you attended any regional or national teaching conferences? If so, list them.

APPENDIX B: PRE-OPEN-ENDED QUESTIONS

1. Middle Tennessee State University randomly selected 20 graduate students and asked them if they were satisfied with the University. Only 60% of the graduate students said they were very satisfied. However, the administration claims that over 80% of all graduate students are very satisfied.

    a. Do you believe the administration?

    b. Can you test their claim?

    c. If so, how would you do so?

2. Describe the following terms in your own words:

    a. Sample

    b. Population

    c. Sampling distribution

    d. Variability

    e. *P*-value

    f. Significance level

    g. Significance

    h. inference

3. How would you describe a hypothesis test to someone who has never heard of one?

4. Publishing scientific papers online is fast, and the papers can be long. Publishing in a paper journal means that the paper will live forever in libraries. The *British Medical Journal* combines the two: it prints short and readable versions, with longer versions available online. Is this OK with authors?
The journal asked a random sample of 104 of its recent authors several questions. One question was "Should the journal continue using this system?" In the sample, 72 said "Yes."
(a) Do the data give good evidence that more than two-thirds (67%) of authors support continuing this system? Carry out a one sample *z* test to answer this question.

(b) Interpret the *P*-value from your test in the context of the problem.

(c) Based on your conclusion from part b, what type of error could you have made, Type I or Type II?

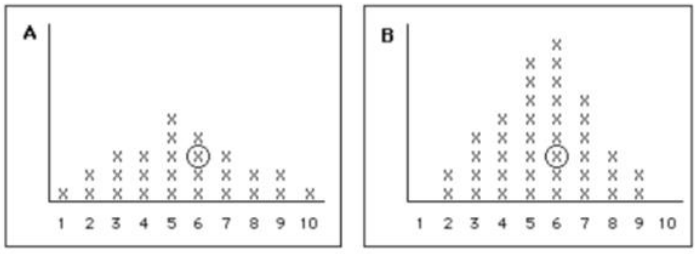
Scenario 10-7
Some researchers have conjectured that stem-pitting disease in peach tree seedlings might be controlled with weed and soil treatment. An experiment was conducted to compare peach tree seedling growth with soil and weeds treated with one of two herbicides. In a field containing 20 seedlings, 10 were randomly selected from throughout the field and assigned to receive Herbicide A. The remaining 10 seedlings were to receive Herbicide B. Soil and weeds for each seedling were treated with the appropriate herbicide, and at the end of the study period, the height (in centimeters) was recorded for each seedling. A box plot of each data set showed no indication of non-Normality. The following results were obtained:

	$\bar{x}$ (cm)	*S* (cm)
Herbicide A	94.5	10
Herbicide B	109.1	9

5. Use Scenario 10-7. Suppose we wished to determine if there tended to be a significant difference in mean height for the seedlings treated with the different herbicides. Do the data provide convincing evidence that the herbicides have a different effect on height?

APPENDIX C: SAMPLE CAOS QUESTIONS (FROM SAMPLING DISTRIBUTION SECTION)

1. Figure A represents the weights for a sample of 26 pebbles, each weighed to the nearest gram. Figure B represents the mean weights of a random sample of 3 pebbles each, with the mean weights rounded to the nearest gram. One value is circled in each distribution. Is there a difference between what is represented by the X circled in A and the X circled in B? Please select the best answer from the list below.



   a. No, in both Figure A and Figure B, the X represents one pebble that weights 6 grams.

   b. Yes, Figure A has a larger range of values than Figure B.

   C. Yes, the X in Figure A is the weight for a single pebble, while the X in Figure B represents the average weight of 3 pebbles.

2. In a geology course, students were learning to use a balance scale to make accurate weighings of rock samples. One student plans to weigh a rock 20 times and then calculate the average of the 20 measurements to estimate her rock's true weight. A second student plans to weigh a rock 5 times and calculate the average of the 5 measurements to estimate his rock's true weight. Which student is more likely to come the closest to the true weight of the rock he or she is weighing?

   A. The student who weighed the rock 20 times.

   b. The student who weighed the rock 5 times.

   c. Both averages would be equally close to the true weight.

APPENDIX D: SEMI-STRUCTURED PRE-INTERVIEW PROTOCOL

1. How do you introduce the concept of hypothesis testing to your students?

2. Why do you think that some students struggle with this topic?

3. Are there any explanations/approaches that you find useful to help your students understand this topic?

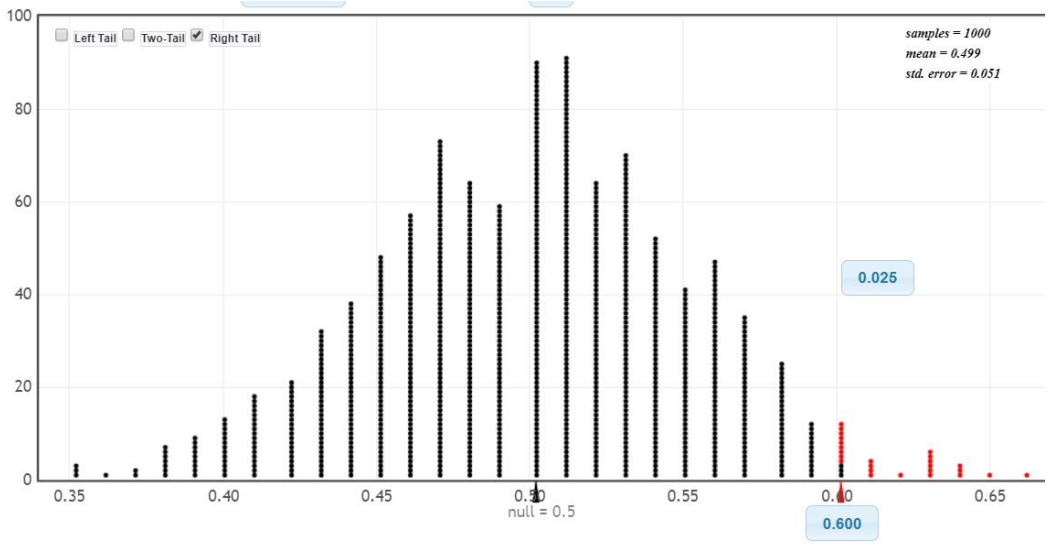
4. What do you think is the pre-requisite knowledge to understand a hypothesis test?

5. What are some common misconceptions students have concerning hypothesis testing?

6. Can you describe when you first learned about hypothesis testing?

7. Ask for elaboration of pre- CAOS and open-ended responses as needed.

8. Assume that sampling procedures are acceptable and that a sample is collected having 60% favoring Pepsi. Argue for or against this conclusion: *This sample suggests that there are more people in the sampled population who prefer Pepsi than prefer Coca Cola*.



APPENDIX E: POST-TASK REFLECTIONS

Post-Task A:

1. Draw a diagram representative of the simulation task that you just completed.
2. In a study, conducted by Yale researchers in 2007, groups of 6-month-olds and 10-month-olds watched a puppet show with neutral wooden figures, where one figure, the climber, was trying to get up a hill. In one scenario, one of the other figures, called the helper, assisted the climber up the hill. In the other scenario, a third figure, called the hinderer, pushed the climber down. Out of the 16 infants in the study, 14 preferred the helper toy. Does this provide statistically significance evidence that the majority of infants prefer the helper toy?
3. Using your diagram, explain how you see the simulation task connected to the traditional hypothesis test problem that you worked in the previous section.

Post-Task B:

1. Create a step by step model/guide that students could use to conduct a hypothesis test using simulations.
2. The *p*-value for a traditional two sample *t*-test for means is .132. Draw an appropriate conclusion.
3. Interpret the approximate *p*-value from the simulation in part c.
4. Interpret the approximate *p*-value from the simulation in part e.

Terminology:

The ***probability*** of an event is the long-run proportion of times the event happens when its random process is repeatedly indefinitely.

The ***p-value*** is the probability that randomness alone would produce data as extreme (or more extreme) as the result obtained in the actual study, assuming the null hypothesis to be true.

A small *p*-value (usually less than .05) indicates that the observed data would be surprising to occur by randomness alone, if the null hypothesis were true. Such a result is said to be ***statistically significant***, and provides evidence against the null.

5. Based on our simulations, make a conclusion given a difference of 6.3 points and justify your conclusion using the above terminology.
6. Explain how you see each step of the traditional approach connected to the simulation approach.

Post-Task C:

1. List the steps for solving any hypothesis test using simulations for inference and write out beside it the corresponding steps for a traditional test.

APPENDIX F: POST OPEN-ENDED QUESTIONS

1. Middle Tennessee State University randomly selected 20 graduate students and asked them if they were satisfied with the University. Only 60% of the graduate students said they were very satisfied. However, the administration claims that over 80% of all graduate students are very satisfied.

   a. Do you believe the administration?
   b. Can you test their claim?
   c. If so, how would you do so?

2. Describe the following terms in your own words:

   a. Sample
   b. Population
   c. Sampling distribution
   d. Variability
   e. *P*-value
   f. Significance
   g. inference

3. How would you describe the logic behind a hypothesis test?

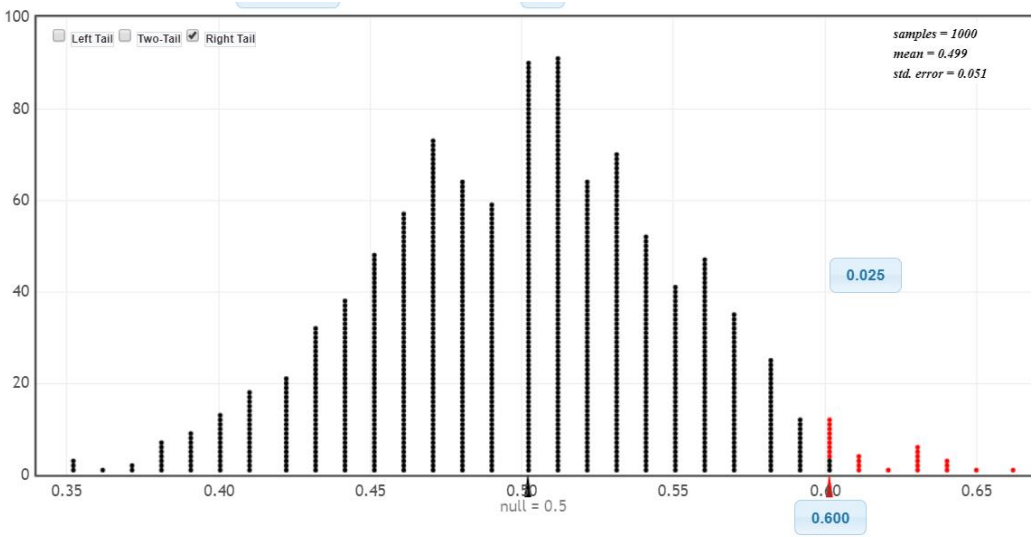
4. What aspects of hypothesis testing did the simulation tasks highlight/emphasize to you?

5. Is there anything about hypothesis testing that these tasks made you think about differently?

6. How do you think these tasks could help students understand hypothesis testing?

APPENDIX G: SEMI-STRUCTURED POST-INTERVIEW PROTOCOL

1. Assume that sampling procedures are acceptable and that a sample is collected having 60% favoring Pepsi. Argue for or against this conclusion: *This sample suggests that there are more people in the sampled population who prefer Pepsi than prefer Coca Cola*.



2. Solve this problem using simulations: Middle Tennessee State University randomly selected 20 graduate students and asked them if they were satisfied with the University. Only 60% of the graduate students said they were very satisfied. However, the administration claims that over 80% of all graduate students are very satisfied. Do you believe the administration's claim?

3. Ask follow-up questions to task responses, reflections, and open-ended response questions.

APPENDIX H: TASK A

Task A: Helper or Hinderer?

In the original study, conducted by Yale researchers in 2007, groups of 6-month-olds and 10-month-olds watched a puppet show with neutral wooden figures, where one figure, the climber, was trying to get up a hill. In one scenario, one of the other figures, called the helper, assisted the climber up the hill. In the other scenario, a third figure, called the hinderer, pushed the climber down.
https://www.livescience.com/7390-babies-judge-character.html

1. If 16 pre-verbal children participated in this study, how many do you think chose the helper toy? Why? What factors do you think might be at play when the children make their choice?

2. How many different possible hypotheses could we make for this situation regarding pre-verbal children and their choice of a toy? What are they?

3. a) If children really do not have a preference for the helper toy, what would be the most likely outcome (# of infants choosing the helper toy) when this study is conducted on 16 infants?

(b) Still assuming that infants show no preference between the helper and hinderer, what kind of results (for number of infants choosing the helper toy) would you not be surprised to see when this study is conducted on 16 infants?
(For example, would 1 out of 16 choosing helper be surprising)?

4. The researchers actually found that _____ of the 16 infants in the study selected the helper toy. If it is REALLY the case that infants show no preference between the helper and hinderer toy do you find the researchers' results surprising? Why or why not?

The key question is, "How surprising is the observed result under the assumption that participants could not really pair the photos correctly (i.e. they were randomly guessing)?" We will call this assumption of randomly guessing the **null hypothesis**. Let's simulate this situation.

5. Design a simulation to represent this experiment assuming that the null hypothesis is true. Carry out three trials of the simulation and record your results below.

_____ _____ _____

6. a) From your results, does it seem like the results obtained by the researcher would be surprising? Explain.

b) Now, we will use technology to simulate this experiment many, many times under the assumption that the null hypothesis is true. Based on this simulation how surprising are the actual results of this study? Explain your reasoning.

c) Based on the results of the simulation, how *likely* would it be to obtain 14 out of the 16 infants choosing the helper toy? Explain.

d) Based on our simulations, what conclusion should the researcher draw? Justify your conclusion.

e) If the actual study had instead found that 9 of the 16 infants chose the helper toy, then what decision should the researchers make based on this result?

APPENDIX I: TASK B

Yellow vs. White Exams

Math teachers often use two different forms of an exam to prevent students from cheating. One trick teachers use is to give the same exam but on two different colors of paper (white and yellow). Some students believe that yellow is a happier, peaceful color compared to the stark white and that they would tend to score better on yellow paper. To investigate this claim, a teacher gave all of her students the same test and randomly chose half the students to take it on white paper and half the students to take it on yellow paper.

1. If 20 students took an exam, 10 on white paper and 10 on yellow paper, how do you think the average score from students who took the exam on yellow paper would compare to average score of the students who took the exam on white paper? Why?

2. How many different hypotheses could we make for this situation regarding the averages of scores of students who take the exam on yellow paper and on white paper? What are they?

3. a) In statistics, we typically subtract the average scores from two groups in order to compare them. If the color of the exam did not affect students' scores, what would be the most likely outcome (difference in the average scores) when this study is conducted with 20 participants?

   b) Still assuming that the color of the exam did not affect students' scores (i.e. students would get the same score regardless of the color of the exam) what kind of results (difference in the average scores) would you not be surprised to see when this study is conducted with 20 participants?

4. For this experiment, the average test score for the yellow paper was _____ and the average test score for the white paper was _____. Therefore, the actual difference in the average scores of students who took the exam on yellow paper compared to students who took the exam on white paper was _____. If it is REALLY the case that the color of the exam doesn't matter, do you find the teacher's result surprising? Why or why not?

What we want to know is, "How surprising is the observed difference in means under the assumption that the color of the exam did not affect the students' scores on the exam (i.e. the students would have gotten the same score no matter which color paper they received)? This assumption is the **NULL HYPOTHESIS.** To test this we will simulate this situation. Think of how you might use randomness to simulate the way that the 20 students were assigned an exam. Note, be sure to simulate the teacher's experiment in which randomness determined which color exam a student receives – not the score they got on the exam.

5. Design a simulation assuming the null hypothesis is true. Carry out three trials of the simulation and record your results below. Be sure to get approval from your teacher before carrying out the simulation.

_____ _____ _____

6. a) From your results, does it seem like the results obtained by the teacher would be surprising? Explain.

b) Now, we will use technology to simulate this experiment many, many times under the assumption that the null hypothesis is true. Based on this simulation how surprising are the actual results of this study? Explain your reasoning.

c) Based on the results of the simulation, how *likely* is a difference of 6.3 or greater? Explain.

d) Based on our simulations, what conclusion should the teacher draw? Justify your conclusion.

e) If the actual study had a difference in means of 10.4 points, then what decision should the researchers make based on this result? Justify your conclusion.

APPENDIX J: TASK C

I'm sad… Let's swim with some dolphins!

Swimming with dolphins can certainly be fun, but is it also therapeutic for patients suffering from clinical depression? To investigate this possibility, researchers recruited 30 subjects aged 18-65 with a clinical diagnosis of mild to moderate depression. Subjects were required to discontinue use of any antidepressant drugs or psychotherapy for four weeks prior to the experiment, and throughout the experiment. These 30 subjects went to an island off the coast of Honduras, where they were randomly assigned to one of two treatment groups. Both groups engaged in the same amount of swimming and snorkeling each day, but one group did so in the presence of bottlenose dolphins and the other group did not. At the end of two weeks, each subjects' level of depression was evaluated, as it had been at the beginning of the study, and it was determined whether they showed "substantial improvement" (reducing their level of depression) by the end of the study (Rossman, 2008).

1. There were 30 participants, and 13 of the 30 participants showed substantial improvement. How many of these 13 improvers do you think were in the "Dolphin Group?" Briefly explain your reasoning.

2. What are all the different hypotheses that we could make for this study?

3. a) In statistics, we typically subtract two numbers in order to compare them. If the null hypothesis is true, what would be the most likely outcome (difference in number of improvers between the dolphin group and the control group)?.

   b) Still assuming that the null hypothesis is true, what kind of results (difference in the number of improvers) would you not be surprised to see when this study is conducted with 30 participants?

4. In the actual study, Antonioli and Reveley found that _____ of 15 subjects in the dolphin therapy group showed substantial improvement, compared to _____ of 15 subjects in the non-dolphin (or the control) group. Complete the table based on these results.

	Dolphin therapy	Control group	Total
Showed substantial improvement			13
Did not show substantial improvement			17
Total	15	15	30

Calculate the difference in the number of improvers between the dolphin group and the control group. Do the data appear to support the claim that dolphin therapy is more effective than swimming alone?

What we want to know is, "How surprising is the observed difference in number of improvers under the assumption that the number of improvers would be about the same for the two groups? This assumption is the **NULL HYPOTHESIS.** To test this we will simulate this situation.

5. Determine a plan for simulating the study. In your simulation plan, how are you representing the participants? How are you representing the improvers and the non-improvers? How will you decide who is in the dolphin and control groups? Be sure to simulate the randomization that the researchers used. How are you recording the results? Describe your plan below and carry out three trials of your simulation.

_____ _____ _____

6. a) From your results, does it seem like the results obtained by the researcher would be surprising? Explain.

b) Now, we will use technology to simulate this experiment many, many times under the assumption that the null hypothesis is true. Based on this simulation how surprising are the actual results of this study? Explain your reasoning.

c) Based on the results of the simulation, how *likely* is a difference of or greater? Explain.

d) Based on our simulations, what conclusion should the researcher draw? Justify your conclusion.

e) If the actual study had a difference in, then what decision should the researchers make based on this result? Justify your conclusion.

APPENDIX K: IRB APPROVAL

IRB
INSTITUTIONAL REVIEW BOARD
Office of Research Compliance,
010A Sam Ingram Building,
2269 Middle Tennessee Blvd
Murfreesboro, TN 37129

MIDDLE TENNESSEE STATE UNIVERSITY

IRBN001 - EXPEDITED PROTOCOL APPROVAL NOTICE

Tuesday, August 22, 2017

Principal Investigator	**Amber Matuszewski** (Student)
Faculty Advisor	Jeremy Strayer
Co-Investigators	NONE
Investigator Email(s)	*alm6p@mtmail.mtsu.edu; jeremy.strayer@mtsu.edu*
Department	Mathematics
Protocol Title	***High school statistics teachers' development of conceptual understanding of hypothesis testing through simulation***
Protocol ID	**18-2005**

Dear Investigator(s),

The above identified research proposal has been reviewed by the MTSU Institutional Review Board (IRB) through the **EXPEDITED** mechanism under 45 CFR 46.110 and 21 CFR 56.110 within the category *(7) Research on individual or group characteristics or behavior* A summary of the IRB action and other particulars in regard to this protocol application is tabulated as shown below:

IRB Action	APPROVED for one year from the date of this notification
Date of expiration	**8/31/2018**
Participant Size	3 (THREE)
Participant Pool	General adult participants (18 years or older)
Exceptions	1. Permitted to compensate ($100) the participants. 2. Recording identifiable contact information is permitted to allow project management, coordination and follow up. 3. Audio recording for data collection is permitted with restrictions.
Restrictions	1. **Mandatory signed informed consent; The PI must provide a signed copy of the informed consent document to each participant.** **2. Participant exclusion criteria MUST be followed as provided in the protocol application.** **3. Identifiable information must be destroyed after data analysis.** **4. Audio recording must be desytroyed once data analysis is over.**
Comments	NONE

This protocol can be continued for up to THREE years (**8/31/2020**) by obtaining a continuation approval prior to **8/31/2018**. Refer to the following schedule to plan your annual project reports and be aware that you may not receive a separate reminder to complete your continuing reviews.

Failure in obtaining an approval for continuation will automatically result in cancellation of this protocol. Moreover, the completion of this study MUST be notified to the Office of Compliance by filing a final report in order to close-out the protocol.

Continuing Review Schedule:

Reporting Period	Requisition Deadline	IRB Comments
First year report	7/31/2018	TO BE COMPLETED
Second year report	7/31/2019	TO BE COMPLETED
Final report	7/31/2020	TO BE COMPLETED

Post-approval Protocol Amendments:

Date	Amendment(s)	IRB Comments
NONE	NONE	NONE

The investigator(s) indicated in this notification should read and abide by all of the post-approval conditions imposed with this approval. Refer to the post-approval guidelines posted in the MTSU IRB's website. Any unanticipated harms to participants or adverse events must be reported to the Office of Compliance at (615) 494-8918 within 48 hours of the incident. Amendments to this protocol must be approved by the IRB. Inclusion of new researchers must also be approved by the Office of Compliance before they begin to work on the project.

All of the research-related records, which include signed consent forms, investigator information and other documents related to the study, must be retained by the PI or the faculty advisor (if the PI is a student) at the secure location mentioned in the protocol application. The data storage must be maintained for at least three (3) years after study completion. Subsequently, the researcher may destroy the data in a manner that maintains confidentiality and anonymity. IRB reserves the right to modify, change or cancel the terms of this letter without prior notice. Be advised that IRB also reserves the right to inspect or audit your records if needed.

Sincerely,

Institutional Review Board
Middle Tennessee State University

Quick Links:

Click here for a detailed list of the post-approval responsibilities.
More information on expedited procedures can be found here.