REASONING ABOUT INFERENCE USING TRADITIONAL AND SIMULATION-BASED INFERENCE MODELS

By

CATHERINE CASE

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2016

For my mom, Dr. Jan Case,
a constant source of inspiration and encouragement

ACKNOWLEDGEMENTS

As I finish this dissertation, I have more thanks than I can put into words. I have the most amazing family: a mom who's a role model in every way, a dad who taught me what it takes to win championships, and a sister who's challenged me and cheered me on since we were kids. I can't even imagine what I would have done without my husband, Adam, who brought me food and copy-edited my writing and never doubted that I would make it through.

Writing a dissertation is tough, and without grad school friends, it may be impossible, but I have the best around. I'm especially grateful for Doug and Steve, who were there from the beginning – my counterparts in the trio with many nicknames, my first colleagues, and some of my dearest friends. I'm forever grateful to Stephanie, Lisa, Carolyn, and Elizabeth, who somehow knew when to encourage me to keep working and when to invite me out for a walk or a cup of coffee. Both in the statistics department and the college of education, there were so many others who supported the writing of this dissertation and got me through the many challenges of graduate school that came before. I couldn't have done it without you.

My sincerest thanks go to my dissertation committee, who have profoundly shaped my work, not only through their generous feedback on this dissertation, but through their dedicated teaching, which gave me new perspectives to study the world. I'm especially grateful to Tim Jacobbe, who invited me into the field of statistics education, introduced me to a community that inspires and supports me, and pushed me to do more than I thought possible. It's a debt of gratitude I can never fully repay.

Finally, I’m thankful for my AP Statistics students at P.K. Yonge Developmental Research School, who generously shared their talents and insights with me and were a consistent reminder of why this is all worth it.

TABLE OF CONTENTS

page

CHAPTER

APPENDIX

LIST OF TABLES

LIST OF FIGURES

Figure page

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

REASONING ABOUT INFERENCE USING TRADITIONAL AND SIMULATION-BASED INFERENCE MODELS

By

Catherine Case

December 2016

Chair: Tim Jacobbe
Major: Curriculum and Instruction

At the recommendation of prominent statistics educators, most notably George Cobb (2007), simulation-based inference methods have begun to replace or complement traditional inference methods in a number of introductory courses, including statistics courses at the high school level (Rossman & Chance, 2014). Developers of curricula that employ simulation-based inference as the primary means of teaching inference have published studies comparing their students' understanding to students in traditional courses (e.g., Garfield et al., 2012; Tintle et al., 2012, 2011). However, these studies, which feature quantitative analysis of student performance on summative assessments, have not provided theory to explain how novices employ the tools and representations of traditional and simulation-based inference models. The existing literature also fails to illuminate student conceptions of inferential topics in courses that employ both traditional and simulation-based methods to introduce the logic of inference.

Traditional inference methods and simulation-based inference methods are two models (and corresponding representational systems) used to express the logic of inference. Using the models and modeling theoretical perspective (Lesh & Doerr, 2003),

this dissertation explores the following central research question: How do students use traditional inference models and simulation-based inference models to understand inference?

The data for this study were collected in an AP Statistics course that employed both traditional and simulation-based inference methods. The data include student responses to targeted formative assessments, exam items, and survey items; daily field notes and journal entries written by the teacher-researcher; and transcripts of individual and group interviews. Data analysis involved a process of systematic coding, following the guidelines for grounded theory provided by Charmaz (2014).

The findings of this study are presented as three scholarly articles. The first examines how students use inferential models, representations, and tools as they reason about a statistical inference task. The second identifies common errors associated with simulation-based inference and characterizes the statistical conceptions underlying those errors. The third illustrates the connections that students make between approaches and offers recommendations for a course that includes both traditional and simulation-based models in instruction.

CHAPTER 1
INTRODUCTION

The discipline of statistics is characterized as a coherent set of tools for dealing with variability in data, and given the “omnipresence of variability” (Cobb & Moore, 1997), these tools are applicable in innumerable contexts. From a scientist studying evolution to a market researcher studying consumer preferences, people from diverse fields need tools to describe how individuals vary within a population. From a researcher conducting a formal analysis to an informed citizen making a decision based on data, people need awareness of variability in measurements; they need familiarity with experiments that involve purposeful variation of certain conditions and control of others. Statistics education should offer opportunities to experience these and other sources of variability and introduce a statistical problem-solving process to make sense of variability in data (American Statistical Association, 2005; Franklin et al., 2007).

In practice, a statistic calculated from a particular sample or experiment is often used to draw inferences about a larger population or an underlying causal relationship. Like data, statistics vary, so statistical inference methods must account for variability in statistics due to random sampling or random assignment. A rich understanding of variability as it relates to statistical inference is essential for students – both those who intend to produce their own statistical analyses and those who will engage with statistics primarily as critical consumers of data-based reports.

In its list of goals for students in introductory statistics courses, the American Statistical Association’s *Guidelines for Assessment and Instruction in Statistics Education (GAISE): College Report* (ASA, 2016) begins with students’ development as *critical consumers* of statistical information. In addition to evaluation of study designs, data descriptions, and data displays, informed consumers should habitually ask whether a reported result – be it a change in a politician's poll numbers or the outcome of a new nutritional study - could have occurred by chance alone. However, "students do not spontaneously raise this possibility" (Konold, 1994, p. 206; Moore, 1990; Pfannkuch, 2005). Absent educative experiences with sampling variability, people instinctively look for deterministic causes rather than consider chance variation (Wild & Pfannkuch, 1999), which may lead to over-interpretation of results.

Future producers of statistics must learn analytical techniques for deciding whether observed results could have plausibly occurred by chance under a given claim or model; in other words, they learn to formally test for statistical significance in different data scenarios. It is important that these students remain grounded in conceptual understanding of inference, recognizing the unifying themes amid the many technical variations (Cobb, 2007).

> Students should not leave their introductory statistics course with the mistaken impression that statistics consists of an unrelated collection of formulas and methods. Rather, students should understand that statistics is a problem-solving and decision-making *process* that is fundamental to scientific inquiry and essential for making sound decisions. (ASA, 2016)

Statistics educators describe diverse inference procedures as ways to model the randomness inherent in study design (Cobb, 2007) and consider observed effects relative to variability (Pfannkuch, 2005). This conceptual foundation is generative as students acquire new inference techniques and adapt these procedures for specific circumstances; further, a strong conceptual foundation prevents misunderstandings that can lead to misuse.

Although significance testing is a ubiquitous data analysis tool, there is evidence to suggest it is misunderstood by many who use it (Nickerson, 2000). In an introduction to the American Statistical Association's statement on the use of p-values, Wasserstein and Lazar (2016) identify misuse and misunderstanding of statistical inference as contributors to a "reproducibility crisis." First, interpretation of p-values is limited; for instance, p-values do not indicate the size or importance of an effect. Second, selective publication of results that pass a specific threshold for statistical significance (e.g., p-value < 0.05) has incentivized poor statistical and scientific practice and caused some to doubt the validity of science itself. These issues "affect not only research, but research funding, journal practices, career advancement, scientific education, public policy, journalism, and law" (Wasserstein & Lazar, 2016). At the same time, terms like "p-hacking" are entering the vernacular, introduced to large audiences by the media, from data-driven journalism websites[1] to late night comedy shows.[2] H.G. Wells is often credited with the prediction, "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.". Already, understanding of statistical inference is important for those who produce and consume statistics in today's world.

Before modern computing power allowed for rapid simulations, introductory statistics courses necessarily relied on traditional methods like z-tests and t-tests to introduce the core logic of inference (Cobb, 2007). Today, a growing number of statistics educators (e.g. Cobb, 2007; Garfield, delMas, & Zieffler, 2012; Lock, Lock, Morgan, Lock, & Lock, 2014; Tintle, VanderStoep, Holmes, Quisenberry, & Swanson,

[1] "Science Isn't Broken," *FiveThirtyEight*, http://fivethirtyeight.com/features/science-isnt-broken/

[2] "Scientific Studies," *Last Week Tonight,:* https://www.youtube.com/watch?v=0Rnq1NpHdmw

2011) are proposing that traditional methods be replaced or supplemented with simulation-based tests. As enrollments in statistics courses grow and simulation-based inference methods gain popularity (ASA, 2016), a research-based understanding of the impact of simulation-based inference becomes necessary. The purpose of this study is to explore how students use traditional and simulation-based inference methods to understand inference in the context of an AP Statistics course.

After introducing statistical concepts relevant to inference, this chapter provides a literature-based rationale for the current study. Although simulation-based inference methods are increasingly prevalent as a means to improve students' inferential reasoning, the impact of these methods is not adequately understood. The models and modeling perspective and a conceptualization of statistical literacy, reasoning, and thinking provide theoretical perspective to address existing gaps in the literature. The chapter concludes with a brief overview of the study and the structure of the dissertation.

Clarification of Statistical Concepts

This section is intended to clarify the statistical concepts central to this study. Working definitions of statistical terms are provided, and an example is presented to illustrate traditional and simulation-based inference methods and highlight the differences between the two.

Statistical Terms

Defined broadly, statistical inference includes four main ideas: significance, estimation, generalizability, and causation (Rossman & Chance, 2014). Significance and estimation are two broad categories of inferential procedures. Significance concerns the strength of the statistical evidence for a particular claim (and against another), and estimation provides an interval of plausible values for a parameter. Generalizability and causation are both related to the scope of inference: To what population can we generalize a statistical conclusion? Is it appropriate to draw conclusions about cause and effect? This dissertation focuses primarily on significance, and the term *inference* should be interpreted narrowly, as it will refer specifically to tests of statistical significance.

Significance tests, also called *hypothesis tests*, are a way to quantify the strength of empirical evidence against a claim – the *null hypothesis* – in favor of another – an *alternative hypothesis* based on theory. Randomized data production (random sampling or random assignment) protects against bias, so the results of a well-designed study provide a valid basis for inference. However, results that appear to support a researcher's claim may be due to chance variability alone; thus, a significance test is necessary to determine whether a "just by chance" explanation is plausible.

The same core logic underlies all significance tests (Cobb, 2007; Garfield et al., 2012; Tintle et al., 2013), and the term *logic of inference* will be used to refer to the following line of reasoning. First, a model is specified to approximate the variability in outcomes that would occur due to randomization alone if the null hypothesis were true. The term *sampling distribution* will refer to this distribution of outcomes, regardless of whether the distribution is developed theoretically or empirically. An observed outcome that was unlikely to occur by chance provides evidence against the hypothesized model, and a *p-value* quantifies the likelihood that the observed outcome occurred by chance. Thus, when the p-value is small, we reject the null model, ruling out a "just by chance" explanation for the observed outcome and concluding in favor of the researchers' claim

(Cobb, 2007; Garfield et al., 2012; Tintle et al., 2013). When we reject the null hypothesis, the results are called *statistically significant*.

Traditionally, statisticians used theoretical probability distributions to model the outcomes that would occur by chance under the null hypothesis. In this study, statistical significance tests based on theoretical distributions (e.g., Normal distribution, *t* distribution, X^2 distribution) will be called *traditional inference methods*. Alternatively, chance outcomes under the null hypothesis can be modeled using *simulations*, which employ physical chance devices (e.g., coins, dice, spinners) or a computer to mimic a random process. Significance tests that use simulations to model the null hypothesis will be called *simulation-based inference methods*. (Elsewhere in the literature, these are sometimes called *randomization-based inference methods*.) Both traditional tests and simulation-based tests are inference methods used by statisticians to account for variability in statistics. However, this study focuses on the use of these methods in instruction, as two ways to model the logic of inference in an introductory statistics course.

Illustrative Example

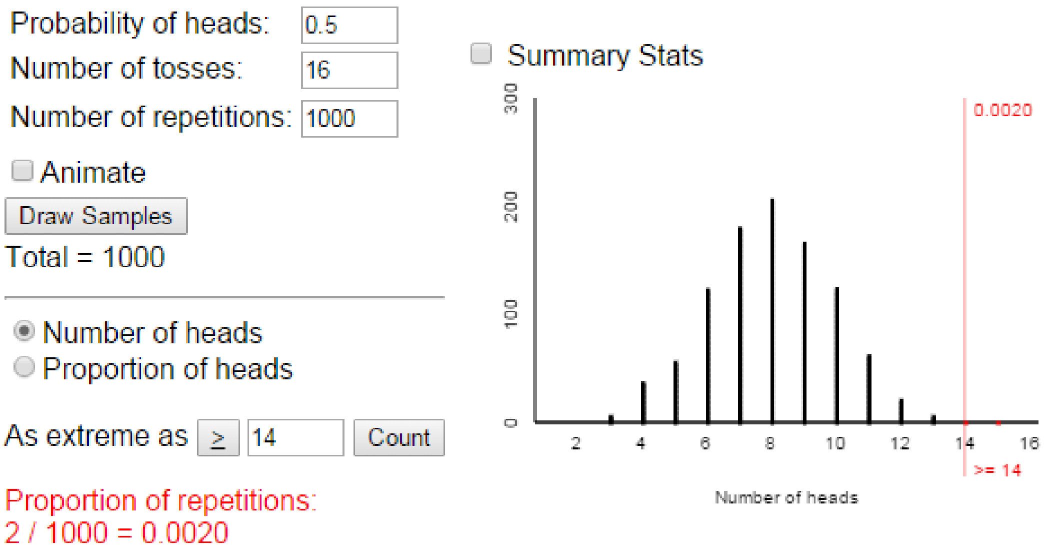
Consider the following example, which necessitates an inference method to determine the statistical significance of experimental results.

> [In] a study reported in *Nature,* researchers investigated whether infants take into account an individual's actions towards others in evaluating that individual as appealing or aversive, perhaps laying the foundation for social interaction. In one component of the study, 10-month-old infants were shown a "climber" character … that could not make it up a hill in two tries. Then they were shown two scenarios for the climber's next try, one where the climber was pushed to the top of the hill by another character ("helper") and one where the climber was pushed back down the hill by another character ("hinderer"). The infant was alternately shown these two scenarios several times. Then the child was presented with the two characters from the video (the helper and the hinderer) and asked to pick

> one to play with. The researchers found that 14 of the 16 infants chose the helper over the hinderer. (Holcomb, Chance, Rossman, Tietjen, & Cobb, 2010, pp. 1–2)

Does this result provide convincing evidence that the infants have a genuine preference for the helper toy or could the result have occurred merely by chance? There are two approaches to answer this question: one based on simulation and one based on theoretical distributions.

Simulation-based inference. Suppose we choose a simulation-based inference method to address this question. If we assume that the infants do not prefer either toy over the other, then their selections can be modeled as a coin flip. For each trial, we flip a coin 16 times to represent the 16 infants who selected a toy in the original study. We can use an applet[3] to simulate data for many trials, recording the number of infants who select the helper each time. The results of this simulation are shown in Figure 1-1.



Probability of heads: 0.5

Number of tosses: 16

Number of repetitions: 1000

□ Animate

Total = 1000

● Number of heads

○ Proportion of heads

As extreme as ≥ 14

Proportion of repetitions: 2 / 1000 = 0.0020

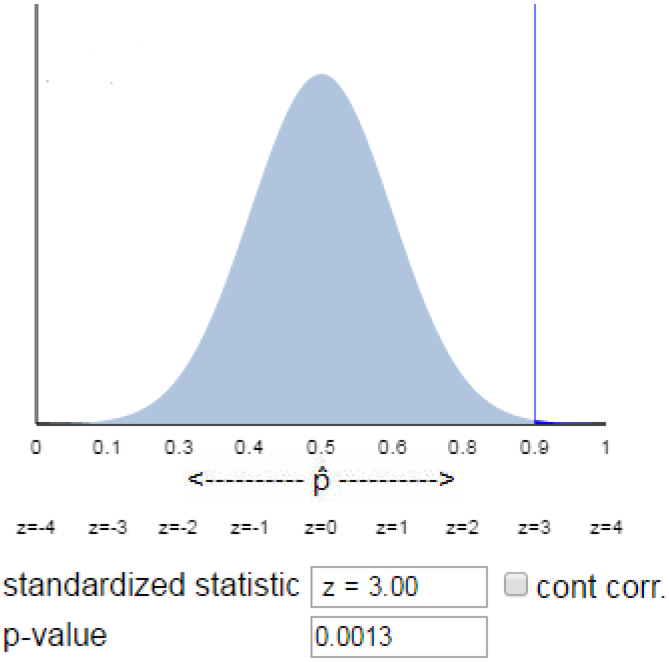
Figure 1-1. Simulation-based test of one proportion using an applet.

[3] Available at http://www.rossmanchance.com/applets/OneProp/OneProp.htm

We evaluate the results by comparing the outcome of the original study to the distribution of outcomes produced by the model. Out of 1000 simulated trials, there were only 2 where heads appeared 14 or more times. That is, if we assume the infants were choosing randomly, the probability of getting a result at least extreme as the observed result (14 or more infants selecting the helper) is about 2 out of 1000; the estimated p-value is 0.002. Because 14 out of 16 infants choosing the helper is very unlikely to occur by chance, we reject the hypothesis that the infants were choosing randomly, and conclude that they have a genuine preference for the helper toy.

Traditional inference. Alternatively, suppose we chose a traditional inference method; a theoretical probability distribution would be used to model the distribution of outcomes that would occur under the null hypothesis. In this case, the binomial distribution could be used to calculate the exact probability of 14 or more babies choosing the helper, assuming that each of the 16 babies chooses independently with probability 0.5. The test would result in a p-value of 0.0021; this p-value is very similar to the one obtained through simulation and has the same interpretation. However, tests based on the binomial distribution are not included in the AP Statistics curriculum and are not taught in many introductory statistics courses.

More commonly, introductory statistics courses use a *z*-test to draw inferences about a single proportion. If certain conditions are met (questionable in this case, because the sample size is small), a sample proportion has a sampling distribution that is approximately Normal with a known standard deviation. This prerequisite knowledge can be used to calculate a standardized test statistic, $z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} = 3$, which suggests that the sample proportion $\hat{p} = \frac{14}{16} = 0.875$ is three standard deviations above the hypothesized proportion of $p_0 = 0.5$. As shown in Figure 1-2, a *z*-statistic has a standard Normal distribution (assuming the null hypothesis is true), which can be used to calculate the probability of obtaining a *z* statistic of 3 or greater.[4] This test would result in a p-value of 0.0013.



standardized statistic z = 3.00 □ cont corr.

p-value 0.0013

Figure 1-2. Traditional test of one proportion using the z distribution.

This brief example illustrates that although the logic of inference is the same for traditional and simulation-based inference methods, the two approaches use different models for the outcomes that are expected under the null hypothesis. Further, the approaches apply different representations (e.g. formulas, graphical displays, notation, and chance devices) and prerequisite knowledge. This dissertation is concerned with how students use the two models and representational systems in courses that employ both to introduce the logic of inference.

[4] The applet used to calculate the p-value is available at http://www.rossmanchance.com/applets/TBIA.html

Theoretical Perspective

This exploration of how students use traditional and simulation-based inference methods will be informed by a *models and modeling perspective*, which conceptualizes learning as model-building, an iterative process in which students invent, extend, and revise constructs (Lesh & Doerr, 2003b). In this study, the term *model* will be defined broadly:

> Models are conceptual systems (consisting of elements, relations, operations, and rules governing interactions) that are expressed using external notation systems, and that are used to construct, describe, or explain the behaviors of other system(s)—perhaps so that the other system can be manipulated or predicted intelligently. (Lesh & Doerr, 2003b, p. 10)

This definition suggests that models have two parts: conceptual systems in the minds of learners and various representational media (Johnson & Lesh, 2003). Thus, models and representations are intimately related, as “the meaning of a model, or conceptual system, tends to be distributed across a variety of interacting representation systems” (Lesh & Doerr, 2000, p. 239). In this dissertation, the term *representation* will be reserved for external systems, such as graphs, equations, and concrete models. In addition to external representations, this study will consider other tools that support statistical inference, including calculators, computer programs, and problem-solving heuristics.

An inclusive definition of the term *model* supports two major research approaches: the constructivist approach focuses on mental models while the pragmatic approach focuses on the functionality of explanatory models and representations (Seel, 2014). The pragmatic approach has been used in subject-matter domains such mathematics to study how external representations mediate understanding by allowing students to simplify complex phenomena, envision the invisible, or identify relationships through analogies (Seel, 2014). This study is situated in that pragmatic tradition.

Central to the models and modeling perspective is the assumption that reality is accessed through models and representations that emphasize different aspects of the underlying system (Lesh & Doerr, 2003b). The discipline of statistics consists of a distinctive set of models and representations of real-world phenomena (Wild & Pfannkuch, 1999), so attention to models and modeling is a powerful lens in this content domain. Lehrer and Schauble (2007, p. 157) argue that “representational change both reflects and instigates new ways of thinking about the data;” thus, it is critical to understand how students interact with the various models and representations used in statistics instruction. Viewed through a models and modeling lens, traditional and simulation-based inference methods can be seen as two models (and corresponding representational systems) used to express the same underlying conceptual system – the logic of inference.

This study also draws on the work of Ben-Zvi and Garfield (2004), who describe teaching and learning of statistics in terms of three broad objectives: statistical literacy, statistical thinking, and statistical reasoning. Though there exist competing definitions for these interrelated constructs, differentiating between the three provides a framework for cataloging objectives and learning outcomes related to statistical inference. Under a restricted definition, statistical literacy describes the ability to organize and read data displays, familiarity with vocabulary and symbols specific to statistics, and knowledge of necessary mathematics content, specifically probability. Statistical reasoning involves making connections between concepts and interpreting statistical information to justify conclusions. Statistical thinking describes thought processes which recognize the need for data and statistical methods to make decisions in a world of omnipresent variability (Ben-Zvi & Garfield, 2004).

Traditional and simulation-based inference methods both require statistical literacy, thinking and, reasoning. However, traditional inference tends to make use of certain graphical representations, vocabulary, notation, and prerequisite knowledge, while simulation-based inference uses others. Thus, demands on statistical literacy may differ between the two methods. For example, traditional inference is based on idealized theoretical sampling distributions, which are typically represented graphically as smooth density curves. Other inscriptions – such as, formal specification of the null and alternative hypotheses and formulas used in calculations – may also be used. In contrast, simulation-based inference involves the construction and use of a model to create a simulated sampling distribution. Since simulated sampling distributions are defined in terms of repeating the randomization process many times, they are often represented graphically using dotplots or histograms instead of idealized curves, and they require less prerequisite knowledge of probability (Cobb, 2007).

If meaning is distributed across various representational media (Lesh & Doerr, 2000), then the differing representations employed in traditional and simulation-based inference may impact students' ability to connect concepts and interpret information – namely, their statistical reasoning. Further, Cobb (2007) suggests that the "technical machinery" of inference – which also differs between the two methods – can obfuscate core statistical ideas. This, in turn, may impact students' statistical thinking. Taken together, the models and modeling framework and Ben-Zvi and Garfield's (2004) conceptualization of statistical literacy, thinking, and reasoning provide theoretical perspective for studying students' understanding of inference.

Teaching for Understanding of Inferential Concepts

A rich understanding of inference is an important outcome of an introductory statistics course. The *GAISE College Report* (ASA, 2016) mentions understanding of statistical inference as a key feature of what it means to be statistically educated. Significance testing is a widely used data analysis method (Nickerson, 2000), and although no introductory statistics course can include all hypothesis tests, "a conceptual understanding of the p-value and statistical significance opens the door to a wide array of statistical procedures that utilize this inferential logic" (Lane-Getaz, 2007, p. 10).

Despite the prevalence of hypothesis testing across disciplines, statistical significance and p-values are commonly misunderstood, not only among college students (Aquilonious & Brenner, 2015; Batanero, 2000; Reaburn, 2014) but also among university faculty (Haller & Krauss, 2002; Mittag & Thompson, 2000) and other professionals who use statistics in their research (Nickerson, 2000). In fact, p-values are so often misinterpreted that some have called for abandoning their use altogether (Nickerson, 2000). Nonetheless, some statistics educators (e.g., Chance & Rossman, 2006) believe that simulations have the potential to develop a deeper conceptual understanding of statistical significance and p-values, and today simulation-based inference methods are increasingly common in introductory statistics courses (ASA, 2016; Rossman & Chance, 2014).

Students' Understanding of Inferential Concepts

As mentioned earlier, p-values are a way to quantify the likelihood of an observed outcome under the null hypothesis to determine statistical significance;

formally, this is expressed as rejecting or failing to reject the null hypothesis. Many students in introductory statistics courses understand p-values as a tool for making decisions about the null hypothesis or a way to quantify the strength of evidence but lack an integrated conceptual understanding of what the p-value represents (Aquilonious & Brenner, 2015; Holcomb, Chance, Rossman, & Cobb, 2010; Taylor & Doehler, 2015). Based on a literature review of ten empirical studies, Lane-Getaz (2007) identified a number of common misconceptions about p-values. Many of these involve confusing relationships between inferential concepts or misinterpreting the p-value as the probability that the hypotheses are true or false.

Chance, delMas, and Garfield (2004, p. 295) attribute poor understanding of inference to “the notoriously difficult, abstract topic of sampling distributions.” Because sampling distributions represent the variability of a statistic in repeated sampling, they invoke prerequisite conceptions of variability, distributions, and sampling (Chance et al., 2004), and require “a multi-tiered scheme” that distinguishes between the population distribution, the distribution of a single sample, and the distribution of statistics calculated from multiple samples (Saldanha & Thompson, 2002). Cobb (2007) compares understanding sampling distributions to understanding derivatives:

> The idea of a sampling distribution is inherently hard for students, in the same way that the idea of a derivative is hard. Both require turning a *process* into a mathematical *object*… Students can understand the *process* of drawing a single random sample and computing a summary number like a mean. But the transition from there to the sampling distribution as the probability distribution each of whose outcomes corresponds to taking-a-sample-and-computing-a-summary-number is … a hard transition. (Cobb, 2007, p. 7)

Further, in traditional inference, changing the setting (e.g., comparing two populations instead of making inferences about one population) or changing the statistic of interest (e.g., comparing medians instead of means) requires substantive, technically difficult changes to the model (Cobb, 2007).

Use of Simulations to Teach Inference

Citing the difficulty of inferential concepts like sampling distributions and p-values, many have proposed the use of simulations to develop statistical concepts (e.g., Chance & Rossman, 2006; Cobb, 2007; delMas, Garfield, & Chance, 1999; Pfannkuch, 2005). Though existing empirical studies demonstrate only modest gains in student understanding (delMas et al., 1999), the use of simulation in statistics instruction is common. Mills (2002) provides examples of computer simulation methods used to develop a number of concepts including the central limit theorem, the t-distribution, confidence intervals, the binomial distribution, regression analysis, sampling distributions, hypothesis testing, and survey sampling.

After using simulation to introduce concepts, most introductory statistics courses use traditional inference methods based on theoretical distributions to carry out inference; however, in recent years, simulation-based inference methods have attracted considerable attention (Rossman & Chance, 2014). Statistics educators have suggested that these methods require less prerequisite knowledge, generalize easily to a large number of settings, incorporate modern computing power in a meaningful way, and support conceptual understanding of inference (Chance & Rossman, 2006; Cobb, 2007; Holcomb, Chance, Rossman, Tietjen, et al., 2010).

These proposed advantages have led to increased use of simulation-based inference in high school and college courses. At the high school level, simulation-based inference is included in curriculum documents, including the Common Core State Standards for Mathematics (CCSSM) (NGACBP & CCSSO, 2010). Further, hundreds of thousands of high school students take AP Statistics every year; though the AP Statistics curriculum primarily uses traditional inference methods, some textbooks incorporate simulation-based inference throughout the course (e.g., Starnes, Tabor, Yates, & Moore, 2013). At the college level, some instructors have incorporated a few activities or modules, while others have completely reconceptualized their courses with simulation-based inference as the cornerstone (Rossman & Chance, 2014).

Statement of the Problem

Simulation-based inference methods have begun to replace or complement traditional inference methods in a number of introductory courses, including statistics courses at the high school level. Though philosophical arguments have been made for these changes, further empirical research is necessary to understand how students use traditional inference models and simulation-based inference models to understand inference. Developers of curricula that employ simulation-based inference as the primary means of teaching inference have published studies comparing their students' understanding to students in traditional courses (Garfield et al., 2012; Tintle, Topliff, Vanderstoep, Holmes, & Swanson, 2012; Tintle et al., 2011). However, these studies, which feature quantitative analysis of student performance on summative assessments, have not provided theory to explain how novices employ the tools and representations of traditional and simulation-based inference models. The existing literature also fails to illuminate student understanding of statistical significance in courses that employ both traditional and simulation-based methods to introduce the logic of inference. Further, although simulation-based inference is included in the CCSSM (NGACBP & CCSSO, 2010) and many high school students study inference in AP Statistics courses, a review of the literature did not reveal any studies of high school students' understanding of inference.

Research Questions

Traditional inference methods and simulation-based inference methods are two models (and corresponding representational systems) used to express the logic of inference. Using the models and modeling theoretical perspective, this study addresses the following central research question:

- How do students use traditional inference models and simulation-based inference models to understand inference?

This broad research interest is further refined in the following sub-questions:

- What conceptions of inferential topics do students hold, and how are these related to commonly occurring student errors?
- What connections do students see between the two models and representational systems?

Overview of Study

Because of its emphasis on inductive data analysis for generation of theory (Charmaz, 2014) and its compatibility with various epistemological paradigms (Taber, 2000), a modified grounded theory methodological approach was used in this study. These methods were informed by a models and modeling perspective (Lesh & Doerr, 2003b), which can be used to study how representations such as mathematical equations, graphs, concrete materials, and notation systems mediate understanding (Seel, 2014). Further, this study employed conceptualization of statistical literacy, reasoning, and thinking (Ben-Zvi & Garfield, 2004) to categorize learning goals and outcomes related to statistical inference.

The data for this study were collected during the second semester of instruction in an AP Statistics course that employs both traditional and simulation-based inference methods. The data include student responses to targeted formative assessments, exam items, and survey items; daily field notes and journal entries written by the teacher-researcher; and transcripts of individual and group interviews. Because grounded theory encourages simultaneous data collection and analysis, the collection of classroom data was guided by the analysis, as the researcher aimed to saturate emerging categories.

Data analysis involved a process of systematic coding, following the guidelines provided by Charmaz (2014). The sensitizing concepts of the models and modeling perspective provided a starting point, but other analytical approaches were also considered. Memos were used to document the process of coding and theory development and to draft descriptions of conceptual categories to be presented in the written report.

Structure of the Dissertation

This dissertation study examines how students use traditional and simulation-based inference in a class that uses both to introduce the logic of inference – a pedagogical approach employed in many classrooms but still underrepresented in the literature. The written report consists of six chapters. This chapter (Chapter 1) presents literature and theoretical frameworks that provide a rationale for the study. Chapter 2 is a formal methods section that details methodology, data collection, and data analysis. Chapter 2 also includes a thorough description of instruction in the AP Statistics class that provided the context for this study. Chapters 3, 4, and 5 are articles written for publication independent of the dissertation document. Chapter 3 examines how students use inferential models, representations, and tools as they reason about a statistical inference task. Chapter 4 identifies common errors associated with simulation-based inference and characterizes the statistical conceptions underlying those errors. Chapter 5 is intended as a resource for practitioners interested in complementing traditional inference with simulation-based methods; this article illustrates the connections that students make between approaches and offers recommendations for a course that includes both traditional and simulation-based models in instruction. Finally, Chapter 6 synthesizes the findings across articles, interpreting the results in relation to the overarching research question; limitations, implications, and directions for future research are also discussed.

CHAPTER 2
METHODS

The purpose of this study is to explore how students use traditional and simulation-based inference methods to understand inference in the context of an AP Statistics course. From a models and modeling theoretical perspective, traditional inference methods and simulation-based inference methods are two models (and corresponding representational systems) used to express the logic of inference. A qualitative study employing modified grounded theory methodology was conducted to address the following central research question:

- How do students use traditional inference models and simulation-based inference models to understand inference?

This broad research interest is further refined in the following sub-questions:

- What conceptions of inferential topics do students hold, and how are these related to commonly occurring student errors?
- What connections do students see between the two models and representational systems?

This chapter details the methodology, context, study design, and limitations of the study. In the first section, a literature-based characterization of grounded theory is provided, accompanied by a rationale for the selection of this methodology. The next section sets the stage for the study by providing a detailed description of instruction in the AP Statistics class in which data were collected. The next section describes the study design, explaining how grounded theory methodology guides data collection and analysis in the context of this study. Finally limitations are presented, defining the scope of the study and acknowledging threats to the study's validity.

Methodology

Because of its emphasis on inductive data analysis for generation of theory (Charmaz, 2014) and its compatibility with various epistemological paradigms (Taber, 2000), a modified grounded theory methodological approach was used in this study. Several defining features characterize this methodological approach (Charmaz, 2014; Corbin & Strauss, 1990; Creswell, 2013). First, data analysis in grounded theory is always inductive and aims to generate theory “grounded” in empirical data. Systematic coding of data relies on constant comparison among data, codes, and categories rather than *a priori* theory. Second, data collection and analysis in grounded theory are carried out simultaneously, and the relationship between the two is expected to be bidirectional. That is, ongoing data collection is informed by the analysis of previously collected data in order to saturate emerging categories. Finally, grounded theory employs memoing as a means of analysis and theory development. Memos are analytical notes that provide a space to develop ideas, plan subsequent data collection, and critically reflect on the research process (Charmaz, 2014).

From its inception, grounded theory represented a marriage of competing research traditions (Charmaz, 2014), and some see grounded theory as a way to build on both positivistic and interpretive paradigms (Taber, 2000). Specifically, Conrad (1982, p. 248) advocates for the use of grounded theory in education research as a means to overcome the “objective-subjective dualism – which has provided the justification for the dominance of ‘quantitative’ over ‘qualitative’ techniques.” The statistical generalizability of quantitative studies stands in contrast to the *analytical generalizability* of qualitative studies, which involve “a reasoned judgment about the extent to which the findings from one study can be used as a guide to what might occur in another situation" (Kvale, 1996, p. 233).

Though the content domain of statistics is largely quantitative and statistics education research has traditionally favored quantitative methodology (Gordon, Reid, & Petocz, 2010), there is no contradiction in using qualitative methods to study the teaching and learning of statistics. Specifically, qualitative research in statistics education is instrumental for "uncovering novel insights and new directions for research" (Kalinowski, Lai, Fidler, & Cumming, 2010, p. 32). Although recent policy documents promote experimental designs and quantitative methods for their presumed objectivity, generalizability, and scientific rigor (Groth, 2010), these studies may not provide sufficient information to inform instruction and policy, particularly if they are conducted before educational interventions have been honed, before measurement tools have been validated, or without consideration of local context variables (Groth, 2010; Kalinowski et al., 2010; Lewis, Perry, & Murata, 2006; Schoenfeld, 2007).

Existing quantitative studies have found that student performance on the multiple-choice CAOS assessment (delMas, Garfield, Ooms, & Chance, 2007) is similar for students who study simulation-based and traditional curricula, with simulation-based curricula linked to modest gains on certain topics including modeling and simulation (Garfield et al., 2012), study design and tests of significance (Tintle et al., 2011). However, these studies have not provided theory to explain how novices employ the tools and representations of traditional and simulation-based inference models. In particular, there are no established theoretical frameworks appropriate for deductive data analysis; thus, grounded theory methods of inductive data analysis are well-suited for early work in this research area. Further, the simultaneous collection and analysis of data in grounded theory grants flexibility to explore new directions, while careful memoing supports rigor.

In grounded theory, discussions of the literature review and theoretical framework are controversial because of grounded theory's rejection of "received theory" (Charmaz, 2014). However, as Dey (1999, p. 251) points out, "There is a difference between an open mind and an empty head." Grounded theorists reject the use of theory to deduce hypotheses before collecting data, but theory inevitably shapes the researcher's worldview (Charmaz, 2014). Thus, a theoretical framework – which might be better called a theoretical perspective – provides "sensitizing concepts" as a starting point, but these concepts are not accepted in the analysis until and unless they "earn [their] way into the theory" (Corbin & Strauss, 1990, p. 7).

This grounded theory exploration of how students use traditional and randomization-based inference methods will be informed by a *models and modeling perspective.* As described in Chapter 1, this perspective conceptualizes learning as modeling-building and assumes that reality is accessed through models and representations that emphasize different aspects of the underlying system (Lesh & Doerr, 2003). Working from a models and modeling perspective, Lehrer and Schauble (2007, p. 157) argue that "representational change both reflects and instigates new ways of thinking about the data." Thus this study examines students' interactions with the models, representations, and tools used for statistical inference, remaining sensitive to the ways these may affect or reflect student thinking.

Complementing Traditional Instruction with Simulation-Based Methods

This study is situated in the context of an AP Statistics class at P.K. Yonge (PKY) Developmental Research School. In addition to the prescribed AP Statistics curriculum, which relies on traditional inference methods (College Board, 2010), the course taught at PKY regularly incorporated simulation-based inference methods. This section describes how the course used simulation-based inference activities as a complement to traditional inference in instruction during the 2015-2016 school year.

Consider the following example, which necessitates an inference method to determine the statistical significance of experimental results.

> In a study reported in the *New England Journal of Medicine*, researchers investigated whether fish oil can help reduce blood pressure. 14 males with high blood pressure were recruited and randomly assigned to one of two treatments. The first treatment was a four-week diet that included fish oil, and the second was a four-week diet that included regular oil. At the end of the four weeks, each volunteer's blood pressure was measured again and the reduction in diastolic blood pressure was recorded. The results of this study are shown below. Note that a negative value means that the subject blood pressure increased. (Starnes et al., 2013, p. 245)

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

This example will be used to illustrate the pedagogy specific to this course and the approach to inference embodied in the AP Statistics curriculum more generally.

Traditional Inference Instruction

The AP Statistics course description includes four major topics (College Board, 2010): data analysis and exploration (20-30%), study design (10-15%), probability and simulation (20-30%), and statistical inference (30-40%). Because traditional inference requires substantial prerequisite knowledge, including knowledge of probability and theoretical sampling distributions, traditional inference is typically taught in the final third of the course (Malone, Gabrosek, Curtiss, & Race, 2010). The course textbook used at PKY, *The Practice of Statistics* (Starnes et al., 2013), adheres to this pattern, covering traditional inference at the end of the year. The AP Statistics curriculum includes nine tests of significance (College Board, 2010): large-sample test for a proportion; large-sample test for a difference between two proportions; test for a mean; test for a difference between two means (unpaired and paired); chi-square test for goodness of fit, homogeneity of proportions, and independence; and test for the slope of a least-squares regression line.

To help students organize statistical problems, *The Practice of Statistics* (Starnes et al., 2013) uses the same four-step process throughout the text. The four-step process, as applied to all significance tests, is as follows:

- State: What *hypotheses* do you want to test, and at what significance level? Define any *parameters* you use.
- Plan: Choose the appropriate inference *method*. Check *conditions*.
- Do: If the conditions are met, perform calculations.
    - Compute the test statistic.
    - Find the ***P*-value**.
- Conclude: *Interpret* the result of your test in the context of the problem. (Starnes et al., 2013, p. 552, emphasis in original)

This four-step process is applied to all traditional tests of significance in the course. In addition to its purpose as an organizational tool, following this process ensures that students earn full credit from graders of the AP Statistics exam.

In the following sections, this four-step process is used to outline how traditional inference instruction experienced by the participants in this study.

State

In this course, a traditional inference task begins with a formal inscription of the null and alternative hypotheses and definition of parameters. We want to perform a test of $H_0: \mu_1 - \mu_2 = 0$ versus $H_A: \mu_1 - \mu_2 > 0$, where μ_1 is the true mean decrease in blood pressure for men like the ones in this study whose diet includes fish oil and μ_2 is the true mean decrease in blood pressure for men like the ones in this study whose diet includes regular oil. At this step, we also set a significance level, typically $\alpha = 0.05$.

Plan

If conditions are met, we will conduct a two-sample t-test for $\mu_1 - \mu_2$. As Cobb (Cobb, 2007, p. 2) points out, evaluating “the fit between model and reality” can be technically complicated, even in this seemingly simple case.

> [Suppose] we really want to be able to compare two groups, two means. So we need the expected value of the difference of two sample means, and the standard deviation of the difference as well. So maybe now we go to the two-sample *t* with pooled estimate of a common variance. At least that *t*-statistic still has a *t*-distribution. But of course it’s not what we really want our students to be using, because it can give the wrong answer when the samples have different standard deviations. So we introduce the variant of the *t*-statistic that has an un-pooled estimate of standard error. But this *t*-statistic no longer has a *t*-distribution, so we have to use an approximation based on a *t*-distribution with different degrees of freedom. (Cobb, 2007, p. 6)

Cobb (2007, p. 8) also disputes the use of a sampling modeling to represent the outcomes of randomized experiments: “Do we want students to leave their brains behind and pretend, as we ourselves apparently pretend, that choosing at random from a large normal population is a good model for randomly assigning treatments?”

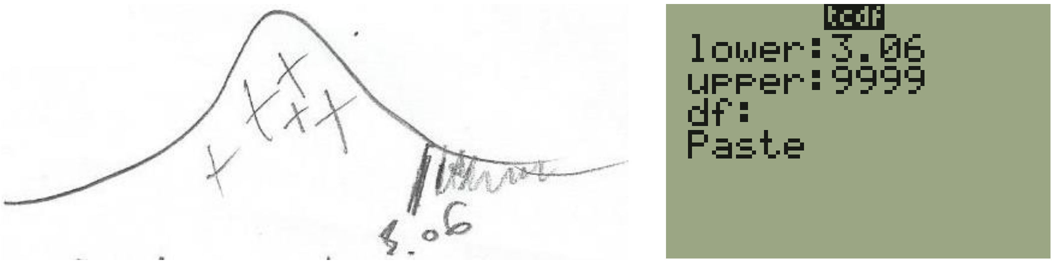
However, this course does not expose students to all the details of theoretical models; the textbook (Starnes et al., 2013) streamlines the process of checking model fit by including lists of conditions for each test. For all tests, students check the *Random* condition; this condition is satisfied since the subjects were randomly assigned to treatments. Next, for all z-test and t-tests, students check the *Normal* condition; the guidelines used to check this condition vary from test-to-test. Because the sample sizes are less than 30 in this example, a t-test is only appropriate if we can reasonably assume that the actual changes in blood pressure are Normally distributed for the two groups. To receive full credit, students must draw a graph to show that they checked the distribution of the sample data. In this example, plots of the data do not show strong skew or outliers, so students would proceed with *t* procedures. Lastly, students check the *Independent* condition. Since these subjects were randomly assigned, the two groups are considered independent; further, one subject's blood pressure should have no effect on another's.

Do

In this course, the calculations for a given hypothesis test are always introduced using formulas for the test statistic. The formula sheet, which students were allowed to use on all exams, includes a general form that expresses the test statistic in terms of the statistic and the hypothesized parameter, but it does not include a complete list of test statistics. In this case,

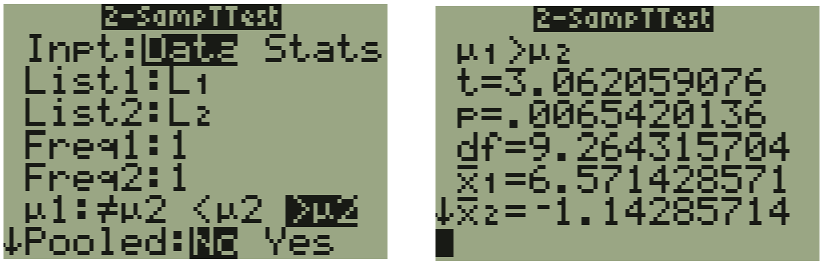
$$t = \frac{statistic - parameter}{st\ dev\ of\ statistic} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(6.57 - (-1.14)) - 0}{\sqrt{\frac{5.86^2}{7} + \frac{3.18^2}{7}}} = 3.06$$

After calculating the test statistic, the p-value is illustrated by sketching a t-distribution and shading the values of *t* more extreme than the one obtained from the original study. If the test statistic and sampling distribution are known, the p-value can be calculated using a cumulative density function in the TI-84 Plus calculator. Figure 2-1 below shows a student-drawn density curve and the calculator function used to find the p-value; the symbols on the curve indicate that it represents a distribution of *t* statistics.



```
tcdf
lower:3.06
upper:9999
df:
Paste
```

Figure 2-1. Calculating a p-value using a cumulative density function.



```
2-SampTTest
Inpt:Data Stats
List1:L1
List2:L2
Freq1:1
Freq2:1
μ1:≠μ2 <μ2 >μ2
↓Pooled:No Yes
```

```
2-SampTTest
μ1>μ2
t=3.062059076
p=.0065420136
df=9.264315704
x̄1=6.571428571
↓x̄2=-1.14285714
```

Figure 2-2. Calculating a p-value using a TI-84 Plus inference function.

For each significance test that is introduced, students first learn to calculate the test statistic using a formula and to find the p-value using a cumulative density function in the calculator. After using these tools a few times, students are introduced to functions on the TI-84 Plus that calculate test statistics and p-values using summary statistics or raw data as inputs, as shown in Figure 2-2.

Two sample t-tests are an exception to the usual pattern. As mentioned by Cobb (2007), there is no simple expression for the degrees of freedom of the *t* distribution used to model the sampling distribution in this context. Thus, students transition to the calculator’s inference functions almost immediately, because “by hand” calculations do not agree with the more complex formulas used by the calculator. Note that the calculator’s inference functions allow students to calculate the test statistic and p-value, as required for the AP Statistics exam, without using formulas or creating visual representations of the sampling distribution.

Conclude

Finally, students must interpret the results of the test in context. Since the theory-based p-value 0.0065 is smaller than the specified significance level, we reject the null hypothesis that the two types of oil had equal effects on the subjects’ blood pressure. If the treatment had no effect, a *t* statistic of 3.06 or larger would be very unlikely. We conclude that on average, fish oil caused larger reductions in blood pressure than regular oil.

To summarize the traditional inference instruction in this course, the final third of the course is devoted to statistical inference (including significance tests and confidence interval estimation). Over a period of several months, students apply the State-Plan-Do-Conclude process to many traditional inference tasks in class, on assignments, and on tests. Following this framework, hypotheses are stated using formal inscriptions. The appropriateness of a particular test is determined by checking a list of conditions. Calculations are introduced through formulas, but later carried out largely by dedicated inference functions in the TI-84 Plus calculator. Conclusions based on the p-value are always interpreted in context.

Simulation-Based Inference Instruction

In addition to traditional inference methods, the course taught at PKY regularly incorporated simulation-based inference methods. Because these methods seem to require less prerequisite knowledge (Cobb, 2007), simulation-based inference can be introduced much earlier in the year, and the AP Statistics course at PKY introduced simulation-based inference on the first day of class.

The use of simulation-based inference to complement traditional instruction was supported by the course textbook, *The Practice of Statistics* (Starnes et al., 2013); however, the course also used activities drawn from other sources. In 2015-2016, the class included fourteen in-class simulation-based inference activities; detailed information about these activities is provided in Appendix E.

To help students recognize the unified modeling process of simulation-based inference, the teacher adopted the 3S Strategy used in the *Introduction to Statistical Investigations* curriculum (Tintle, et al., 2013), applying the same approach to each inference task:

> 1. Statistic: Compute the statistic from the observed sample data.
> 2. Simulate: Identify a “by chance alone” explanation for the data. Repeatedly simulate values of the statistic that could have happened when the chance model is true.
> 3. Strength of Evidence: Consider whether the value of the observed statistic from the research study is unlikely to occur if the chance model is true. If we decide the observed statistic is unlikely to occur by chance alone, then we can conclude that the observed data provide strong evidence against the plausibility of the chance model… (Tintle, et al., 2013)

This three-step process is applied to all simulation-based tests of significance. In the following sections, the process is used to outline how simulation-based inference is taught in this course.

Statistic

The data from the original experiment can be summarized by a difference of sample means: $\bar{x}_1 - \bar{x}_2 = 6.57 - (-1.14) = 7.14$. Unlike traditional methods, this approach does not require calculation of a standardized test statistic. There are two possible explanations for the observed difference of sample means: The first is a "by chance alone" explanation; it is possible that the treatments had no effect, and the difference between groups is due to random assignment. The second is that fish oil caused larger reductions in blood pressure than regular oil, on average.

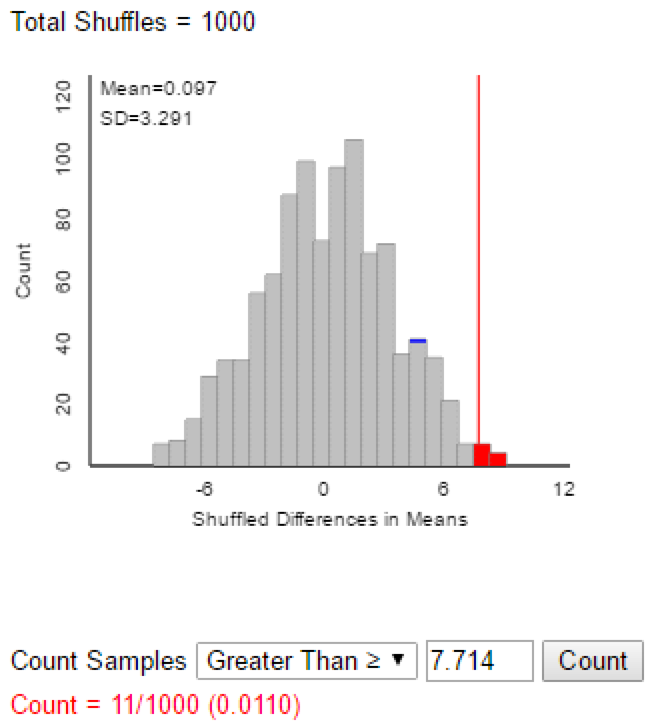
Simulate

Random assignment to groups can be modeled using cards. First we write the improvement scores on blank cards. Then cards from both groups are shuffled together, assuming that the subjects' improvement scores were determined in advance by factors unrelated to treatment. Then the cards are dealt into two groups to mimic random assignment, and the difference in means for the two groups is recorded. We can use an applet[1] to simulate data for many trials. The results of this simulation are shown in Figure 2-3.

Students may engage with the simulation in a number of ways. First, they may be asked to physically carry out the simulation using cards; students may record their own simulated statistics on a class dotplot that can be used to determine the strength of evidence. Second, students may use the applet on their tablets, phones, or computers. Third, students may be asked to design an appropriate simulation. These modes of engagement are not mutually exclusive, and this course used each of them in varying

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

combinations over the course of the year. The mode of student participation in the simulation step largely determines how much time is spent on these activities.



Total Shuffles = 1000

Mean=0.097
SD=3.291

Count Samples Greater Than ≥ ▼ 7.714 Count

Count = 11/1000 (0.0110)

Figure 2-3. Applet to simulate random assignment

Strength of evidence

We evaluate the strength of evidence by comparing the outcome of the original study to the distribution of outcomes produced by the model. In the simulation shown in the figure, a difference of means of 7.714 or larger occurred in 11 out of 1000 trials – an estimated p-value of 0.0011. Because a difference of 7.714 would be very unlikely to occur by random assignment if the treatment had no effect, we reject the “by chance alone” explanation, and conclude that fish oil caused larger reductions in blood pressure than regular oil, on average.

To summarize, this course incorporated simulation-based inference throughout the year, beginning on the first day of class, and added traditional inference later in the second semester. In total, the course included fourteen in-class experiences with simulation-based inference, including multiple opportunities for groups of students to design and carry out simulations using physical chance devices and applets. In the second semester, simulation-based inference activities were included as part of every chapter, so students used simulations to model a variety of study designs. However, because the AP Statistics course description emphasizes traditional inference, students had considerably more experience with traditional, theory-based methods by the end of the school year.

Lastly, this course intended to make the connections between simulation-based inference and traditional inference explicit. For example, traditional inference was first introduced as a modification to the 3S Strategy, where the simulation step was replaced by use of a theoretical sampling distribution. Strategies used to promote connections are described in more detail in Chapter 5. Further, assessments prompted students to reflect on the connections between approaches; assessment tasks are described later in this chapter.

Study Design

Context

This study is situated in the context of an AP Statistics class at P.K. Yonge (PKY) Developmental Research School. PKY is a public school district affiliated with the University of Florida that serves students in grades K-12. The admissions policy of PKY aims to create a student body that mirrors the demographics of the state of Florida for five admission categories: gender, race/ethnic origin, family income, exceptional student status, and academic achievement level. PKY uses block scheduling, and the AP Statistics course meets three times per week throughout the academic year.

Data for this study were collected in two phases. First, a pilot study was conducted at the end of the 2013-2014 academic year. Building on the results of the pilot study, a second phase of data collection was conducted throughout the spring semester of the 2015-2016 academic year. In 2013-2014 – the first year that AP Statistics was offered at PKY – the course was taught by a four-person instructional team composed of a high school mathematics teacher and three Ph.D. students studying statistics education, including the author. Twenty-one students from grades 10 – 12 completed the class and took the AP exam. In 2015-2016, the course was taught by a single instructor - the author. Eleven students from grades 10 – 12 completed the class and took the AP exam.

Participants

Sampling in grounded theory aims to gather enough information to support theory development (Creswell, 2013): "in grounded theory, representativeness of concepts, not of persons, is critical" (Corbin & Strauss, 1990, p. 9). Consequently, the necessary number of participants required cannot be specified concretely, as it depends on factors including topic and research purpose (Charmaz, 2014). In the field of statistics education, qualitative studies intended to characterize students' statistical literacy, reasoning, and thinking often involve in-depth analysis of observations or interviews with a relatively small number of participants (e.g., Aspinwall & Tarr, 2001; Stohl & Tarr, 2002; Wild & Pfannkuch, 1999; Zieffler, delMas, Garfield, & Brown, 2014)., 2014). Further, grounded theory maintains that data collection and data analysis are intertwined; the principle of *theoretical sampling* suggests that additional participants and incidents should be selected for their relevance to the emergent theory until the categories become saturated (Creswell, 2013). Thus, grounded theory methodological principles support multi-phase data collection.

At the pilot study phase, seven students were selected to represent a range of statistical understanding, as judged by their course grades. This decision was based on the assumption that students with varying achievement levels may use traditional and simulation-based inference models differently, thus contributing different information to the developing theory. Additionally, the sample of seven students reflected the demographic diversity of the class, in terms of sex, race, and grade level. A demographic description of participants in the pilot study is given in Table 2-1.

Table 2-1. Demographic description of participants in pilot study.

Sex		Race		Grade level	
Male	3	White	4	10th	4
Female	4	Black	1	11th	3
		Asian	1	12th	2
		Hispanic	1		

Creswell (2013, p. 86) describes grounded theory data collection and analysis as a “zigzag” process: “out to the field to gather information, into the office to analyze the data, back to the field to gather more information… and so forth.” In the second phase of the study, data were collected to answer questions raised by preliminary analysis of the pilot data. Data were collected from all students enrolled in the course in an effort to saturate the emerging categories. However, data collection was guided by concurrent data analysis, so data collection was not identical for each participant. For example, if a student demonstrated a particular conception on a formative assessment, a follow-up interview question might be asked of that students that was not asked of others. A demographic description of participants in the second phase is given in Table 2-2.

Table 2-2. Demographic description of participants in second phase.

Sex		Race		Grade level	
Male	3	White	7	10th	2
Female	8	Black	0	11th	3
		Asian	1	12th	6
		Hispanic	3		

Before beginning this study, permission was granted by the institutional review board (UFIRB-02 for social and behavioral research). The UFIRB based their decision on a submitted protocol that described the purpose of the study, the research methodology, the potential benefits and risks, and the recruitment plan. According to the approved plan, students assented to participate in the study by signing an informed consent document. Because some students were under 18, the informed consent document also required a parent's signature to indicate consent for the minor to participate. In addition to the risks and benefits of participation in the study, the informed consent document explained that participation is voluntary and that students who choose not to participate at any point will face no consequences (academic or social). The informed consent letter is included in Appendix B.

Role of the Researcher

As implied above, the author played a dual role in this study: teacher and researcher. Creswell (2013) suggests that researchers who are deeply engaged as participants may benefit by establishing greater rapport and gaining insider views. In this study, the dual role provided ample opportunity for data collection and intimate knowledge of the instructional context. However, engagement in classroom activity can also distract from research activity (Creswell, 2013). Efforts to overcome this challenge are described in the data collection section.

Data Collection

As described above, the data for this study were collected in the context of an AP Statistics course that employed both traditional and simulation-based inference methods. The data include student responses to targeted formative assessments, exam items, and survey items; daily field notes and journal entries written by the teacher-researcher; and transcripts of individual and group interviews. Data for the pilot study consisted of individual interviews only. A timeline for data collection in this study is provided in Table A-1 in Appendix A.

In contrast to the positivist leanings of early grounded theorists, this study recognizes that data are not merely collected, but co-constructed by the researcher and the participants:

> In the original grounded theory texts, Glaser and Strauss talk about discovering theory as emerging from data separate from the scientific observer. Unlike their position, I assume that neither data nor theories are discovered either as given in the data or the analysis. Rather, we are part of the world we study, the data we collect, and the analyses we produce. We *construct* our grounded theories through our past and present involvements and interactions with people, perspectives, and research practices (Charmaz, 2014, p. 17)

Thus, these descriptions of data sources acknowledge how the data were elicited, with particular attention given to the role of the researcher and the influence of concurrent analysis.

Student work

Formative assessments and exam items were intended to assess students' emergent understanding of inference and associated concepts, such as sampling distributions and p-values. Additionally, some items prompted students to reflect on their use of models and representations or draw connections between inferential concepts. That is, student work informs all three research questions. Appendix D includes a complete list of assessment items used to elicit student work for this study. These items are indexed by date administered and by the associated research question. Some items focus on a single research question while others prompt responses that may inform multiple research questions. Consider the following illustrative examples of assessment tasks.

In one class period, simulation-based inference was used to test a hypothesis about the difference between two proportions. At the end of class, students were asked to write a few sentences about the role of the simulation. Additional prompts were provided to help them get started: Why is [the simulation] necessary? What does the simulated distribution represent? How do we use the simulation to make a decision about the hypotheses? In the next class period, traditional inference was used to test a hypothesis about the difference between two proportions. At the end of class, students were asked to write a few sentences about the role of the sampling distribution, with prompts that mirrored the ones given on the previous assessment. Student responses to these formative assessments illuminated how students use traditional and simulation-based inference models (research question 1). Further, responses revealed students' conceptions of inferential topics, particularly sampling distributions (research question

2). Finally, comparing responses across these items suggested connections perceived between the two approaches (research question 3).

Some items assessed connections more directly. For example, on the day that z-tests were introduced, students were asked to explain how a z-test is similar to the 3S strategy and how it is different. Consistent with a models and modeling perspective, other items highlighted tools and representations. For example, one presented students with two graphical representations: a dotplot representing a simulated sampling distribution and a smooth density curve representing a theoretical sampling distribution. Students were asked to compare and contrast the two graphical representations. How are they similar? How are they different? Responses to questions like these illuminate the connections students see between the two models and representational systems (research question 3).

Formative items like the illustrative examples above were not graded, whether they were administered in class (e.g., as an exit ticket) or as part of a chapter exam. The body of data also includes student responses to graded assessments items that ask students to carry out inference procedures. Lastly, the data includes responses to open-ended assessments that students completed as a group. For example, students were asked to design a simulation that could be used to test hypotheses in a particular context.

Because grounded theory encourages simultaneous data collection and analysis, the collection of student work was guided by the analysis, as the researcher aimed to saturate emerging categories. For example, analysis of pilot study data piqued the researcher’s interest in the connections that students make between traditional and simulation-based models and students' conceptions of specific inferential topics, like sampling distributions and p-values. Thus, many early assessment items were targeted at the second and third research questions. However, analysis of assessments and journal entries suggested that open-ended modeling tasks also had the potential to be thought-revealing. In particular, when students designed models and used tools with little teacher intervention, they employed unexpected strategies and revealed their thinking about the larger modeling process. Thus, several assessments administered late in the year are more open-ended design tasks.

Further, tasks focused on conceptions and connections became increasingly targeted as patterns arose in the on-going analysis. Compare the first and last assessment items, both aimed at the second research question. The first simply asks, "In your own words, what is a sampling distribution?" The last asks (among other questions), whether students are surprised that the average of the simulated slopes is near 0; this question was intended to assess a specific conception – the assumption of the null hypothesis in a simulation.

Teacher reflections

Student work as a data source has limitations, because it only captures students' written inscriptions. During class, the teacher-researcher may observe other modes of modeling and representation. For example, the language that students use to describe models during a class discussion may differ from the language they use on a written assessment at the end of the lesson. Thus, data collection also included daily teacher reflections, beginning at the point in the course when traditional inference was introduced.

Immediately after each lesson, the teacher-researcher wrote a journal entry to record her observations of students' statistical literacy, reasoning, and thinking. These journal entries were based on brief, informal field notes taken during the lesson. First, the teacher-researcher briefly described the day's lesson to provide context for other data collected. Then, she used the research questions as prompts to capture observations of classroom activity that informed the emerging theory. Guided by the sensitizing concepts of the models and modeling framework, the teacher-research carefully attended to students' use of mathematical equations, graphs, concrete materials, notation systems, and specialized language.

The teacher reflections played an important part in the larger data collection process. First, these journal entries complemented students' written work with information about the context and the process. For example, when groups worked on a modeling task, they only submitted their final models; however, the teacher was able to observe the process by which they proposed, rejected, and/or revised models. Second, journal entries often informed future data collection. Consider again the example of students working on modeling tasks in groups. The teacher's journal entry written after one such activity included the following quote:

> From a methodological standpoint, it is interesting to see how students interact with each other in groups without intervention from me. It gives me a better sense of their language and the approaches that they use. However, it was very difficult to monitor and record all that was going on. What I overheard from their conversations or learned by asking probing questions was at least as interesting as what they wrote on paper. Their written responses capture their best "answers" not the thinking that led to those answers.

This real-time observation informed future data collection, particularly the design of the group interviews.

Individual interviews

The data for this study include task-based individual interviews (Maher & Sigley, 2014) with two cohorts of AP Statistics students: Seven students from the 2013-2014 cohort were interviewed in May 2014, and seven students from 2015-2016 cohort were interviewed in May 2016. The individual interviews, which lasted between 20 and 45 minutes, were conducted during the school day in the weeks immediately following the AP Statistics exam. These interviews were audio-recorded and transcribed, and students' written work was collected. The interview tasks prompted students to conduct hypothesis tests to draw conclusions about the results of research studies. The research studies used as inference tasks in the interviews are included in Appendix D.

In the individual interviews, students were asked to apply two different inference methods – a traditional test and a simulation-based test – to a single given context. All tools necessary to carry out the two approaches were provided to students; these included chance devices (coins, dice, cards, etc.), computer applets, formula sheets, and graphing calculators with statistical functions. As students worked, they were encouraged to think aloud and provide any relevant visual representations. Students were also asked to write an interpretation of the p-value as a probability in context after completing the first hypothesis test; they were given an opportunity to revise their interpretation at the end of the interview.

After carrying out both approaches, students were asked to compare and contrast the two approaches and describe any connections they saw between them. After the students' initial responses, the interviewer directed their attention to specific parts of their work and probed for other connections. In particular, students were asked to comment on the similarities and differences between the two approaches in terms of the claims/hypotheses being tested, the conditions necessary to perform the test, the calculations, the representations of the sampling distribution, and the p-values. The individual interview protocol is included in Appendix D.

Schoenfeld (1985) points out that verbal data collected through “out-loud” problem-solving protocols are affected by a number of variables, and the individual interviews used in this study offer specific strengths and limitations. First, single-person protocols offer insight into the knowledge of an individual student, but they may elicit “unnatural” responses, if students feel uneasy receiving individual attention in the interview environment or feel pressure to “produce something mathematical for the researcher” (Schoenfeld, 1985, p. 178). Second, asking probing question about why choices were made may impact students’ behavior: “The students may begin to reflect on those choices while working on the given task, and behave from that point on in a manner very different than he or she would otherwise have behaved” (Schoenfeld, 1985, p. 175). To address some of these limitations, this study also included multi-person protocols, as described below.

Group interviews

All eleven students from the 2015-2016 cohort were invited to participate in group interviews. These interviews were designed to investigate how students use inferential models, representations, and tools while working in groups. Ten students were recorded as they completed statistical inference tasks in pairs. (One student was unable to participate because of absence.) The interviews, which lasted between 25 and 45 minutes, were conducted during the school day in the weeks immediately following the AP Statistics exam. The interviews were audio-recorded and transcribed, and the students’ use of an applet for simulation-based inference was recorded using a screen-capturing tool. Other than the presence of recording devices, this activity was very similar to others students had completed in class.

Similar to the structure of the individual interviews, pairs of students were asked to apply two different inference methods – a traditional test and a simulation-based test – to a single given context. The inference task, which asked students to draw conclusions from an experiment testing response bias, is included in Appendix D. At the beginning of the interview, students were made aware of all available tools – a formula sheet, a graphing calculator with statistical functions, an extensive collection of chance devices (coins, dice, cards, etc.), and a computer applet. However, no instructions were given about whether students should use a traditional or simulation-based approach. After drawing a conclusion, students were asked to consider the alternative approach; e.g. students who initially carried out a traditional test were asked to carry out a simulation-based test.

Unlike the individual interviews, students were not asked to “think aloud” and the interviewer largely refrained from asking questions. The interviewer intervened in the students’ problem-solving process as little as possible, but when students struggled to move forward, the interviewer sometimes prompted them with questions such as, “What are you trying to represent?” or “Would it help you to see the applet?” These were intended to reflect the kinds of prompts students would hear if they asked for help from the classroom teacher. The group interview protocol is included in Appendix D.

A two-person protocol was included in this study for several reasons. First, the group interview data were intended to approximate how students reason about inference in a natural classroom setting, and in class, students typically worked on inference tasks in groups. In particular, classmates who had worked together on similar tasks in class were paired together to help students feel comfortable in the interview environment. In addition to realism and student comfort, Schoenfeld (1985) suggests that interviews with multiple students produce rich data for investigating students' problem-solving processes. Multi-person protocols ease the pressure to "produce something mathematical for the researcher" (Schoenfeld, 1985, p. 178), and discussions among students makes the reasoning behind their decisions more visible (Schoenfeld, 1985).

Schoenfeld (1985, p. 174) reminds us, "Any framework for gathering and analyzing verbal data will illuminate certain aspects of cognitive processes and obscure others: There are trade-offs between structured and unstructured interviews, single or group protocols, etc." Thus, the verbal data collected through individual and group interviews were triangulated with the other data sources.

Confidentiality

Student identity will be kept confidential to the extent provided by law. Specifically, all student information has been assigned a pseudonym. Student names will not be used in any report. Hard copies of data are stored in a locked file cabinet and digital copies stored on a password protected hard drive.

Data Analysis

For this study, the primary goal of the data analysis was to explore how students use the two models and their representational systems as they reason about statistical inference. Data analysis consisted of a process of systematic coding in multiple phases, according to the guidelines for grounded theory presented by Charmaz (2014).

Initial coding

In the initial coding phase, each segment of data in the student work, teacher reflections, and interview transcripts was assigned a concrete and descriptive code to reflect the students' actions. For example, the group interview with Libby and William, discussed at length in Chapter 4, received hundreds of initial codes including the following:

- Proposing a chi-square test
- Reading calculator output
- Deciding to reject the null based on a rule
- Using cards to represent yes/no
- Explaining how the model accounts for treatment groups
- Questioning the validity of using data from one sample
- Suggesting that the observed sample could be an outlier

In keeping with the recommendations put forth by Charmaz (2014), these initial codes are short and precise. They are coded as gerunds in order to focus on students' actions and stay close to the data, encouraging the researcher to "begin analysis from their perspective" (Charmaz, 2014, p. 121).

Focused coding

After comparing these initial codes to the data and looking for patterns in the codes across interviews, the researcher inductively constructed focused codes to "sift, sort, synthesize, and analyze" the large amounts of data (Charmaz, 2014, p. 138). Unlike initial codes, focused codes never retained contextual information specific to the task. Thus, focused codes facilitated comparisons across tasks, across students, and across inferential approaches.

In this study, focused codes reflect the "sensitizing concepts" of the models and modeling perspective, as well as the researcher's familiarity with descriptions of inference by other statistics educators. However, these concepts were not accepted into the analysis until they could be substantiated in the data (Charmaz, 2014; Corbin & Strauss, 1990). In some cases, using the language of modeling, tools, and representations facilitated comparisons. For example, consider two initial codes: checking conditions for a chi-square test and discussing the inadequacies of a coin for modeling yes/no responses in a population where the two responses are not equally likely. At first glance, these do not appear similar, but from a modeling perspective, both are instances of evaluating model fit.

Ultimately, the focused codes were grouped into categories, the largest and most salient of which are presented in Chapters 3, 4, and 5. More specific information about data analysis is presented in the methods section of each chapter.

Memoing

From the earliest coding sessions to the final presentation of results, “writing theoretical memos is an integral part of doing grounded theory” (Corbin & Strauss, 1990, p. 10). Early in the research process, memoing provides a spontaneous, informal way of capturing ideas that will be refined later. Successive memos become more abstract and theoretical as categories are developed (Charmaz, 2014). In this study, memos were used to document the process of data collection and coding and to draft descriptions of conceptual categories to be presented in the written report.

Specifically, grounded theory calls for constant comparisons among data, codes, and categories rather than *a priori* theory; memos provide a space to refine these comparisons. In this study, memos were used to describe comparisons at various phases in the analysis, including the following:

- Comparison of responses on a common task across individuals

- Comparison of data collected from a single individual across tasks
- Comparison of modeling codes across traditional and simulation-based approaches
- Comparison of conception codes across traditional and simulation-based approaches
- Comparison of one common error to another
- Comparison of data coded as an error to data that provide context for the error

Thus, memoing provided space to develop codes, explore connections, and identify conceptual areas where further data collection was necessary.

Following the guidelines of the grounded theory methodological approach and diligently documenting the interrelated processes of data collection and data analysis establish rigor in this qualitative study. Intermittently, progress was compared to the criteria put forth by Charmaz (2014) for grounded theory studies; these include credibility, originality, resonance, and usefulness.

Discussion of Perspective and Scope

In contrast to the statistical generalizability of quantitative studies, this qualitative study aims for *analytical generalizability* (Kvale, 1996). Even in the field of statistics education, which has traditionally favored quantitative methodology (Gordon, Reid, & Petocz, 2010), there have been calls for qualitative research that produces "vivid descriptions that … represent a researcher's well-formulated perspective rather than an objective reality that cuts across a tightly specified range of context" (Groth, 2010). Accordingly, this dissertation describes how students used traditional and simulation-based inference to understand inference in the context of an AP Statistics class; the study's conclusions are supported by data co-constructed by the teacher-researcher and the participating students. This section provides recommendations for how the results might be interpreted in light of the study's design and epistemological position.

First, all data were collected in AP Statistics classes taught by a single teacher-researcher at a single school. It is not reasonable to assume that the results of the study will generalize to all introductory statistics courses that complement traditional inference with simulation-based inference. Factors including the teacher, the textbook, the school environment, and individual student traits likely affected how the participants in this study used traditional and simulation-based inference models to understand inference. In order to help the reader "discover the extent to which the theory does apply and where it has to be qualified for the new situations" (Corbin & Strauss, 1990, p. 15), this dissertation includes thorough descriptions of the context and participants to "situate the sample" (Kalinowski, et al., 2010).

Second, the study provides a naturalistic description of the use of models, tools, and representations in a single classroom environment; it is not an experimental design. Thus, the study does not provide a basis for comparing the effectiveness of traditional and simulation-based inference, as all students were exposed to both approaches in instruction. Further, the study does not provide a basis for comparing simulation-based inference with other possible uses of class time, such as additional practice with traditional methods.

Lastly, the research process in this study is subjective. This limitation is routinely acknowledged for qualitative studies, although Groth (2007) points out that all forms of research involve subjective decisions that largely determine the final conclusions and interpretations that can be drawn. In contrast to the positivist leanings of early grounded theorists, this study acknowledges the subjectivity of both data collection and analysis. Recognizing that data are not merely collected but co-constructed by the researcher and the participants, this dissertation described how the data were elicited with attention to the role of the researcher and the influence of concurrent analysis. Specifically, three facets of the study's subjectivity merit mention.

First, the researcher's dual role as teacher necessarily influenced the study. The teacher-researcher chose to use simulation-based inference in her class, which suggests a certain pre-existing belief in the potential of these methods. Second, the researcher was familiar with views of inference shared among statisticians and statistics educators before she began this study. This lens of disciplinary knowledge likely influenced the researcher's perceptions of the students' modeling practices. Third, incidents in the data were coded by a single person, so there is no indication of inter-rater reliability. This highlights the importance of the teacher-researcher's perspective in constructing the final analysis. Thorough documentation of data analysis and clear, detailed examples "invite the reader to appraise [the researcher's] interpretations and think about other ways the data could have been interpreted" (Kalinowski et al., 2010, p. 30).

Summary

The overarching goal of this dissertation is to explore how students use traditional inference models and simulation-based inference models to understand inference. Because of its emphasis on inductive data analysis for generation of theory and its compatibility with various epistemological paradigms, a grounded theory methodological approach is appropriate for this study. Further, the systematic procedures for data collection and analysis outlined in the grounded theory literature address standards of rigor and support credibility of the study’s results. Currently, the pedagogical approach of complementing traditional inference with simulation-based methods is common in introductory statistics classes but underrepresented in the research literature. In particular, this approach has not been studied in AP Statistics courses at the high school level, and there is no theory to explain how novices use traditional and simulation-based inference models in courses that employ both to introduce statistical inference. Thus, this study represents a useful contribution to the field of statistics education.

CHAPTER 3
STATISTICAL MODELING AS A THOUGHT-REVEALING ACTIVITY

Introduction to Inference Models

Statistical inference incorporates knowledge of both data analysis and probability. More specifically, the logic of inference connects data and chance via the relationship between empirical outcomes and hypothesized models. Introductory statistics courses teach statistical inference using two distinct approaches. *Traditional inference methods* use theoretical probability distributions (e.g., Normal distribution, *t* distribution, X^2 distribution) to model the outcomes that would occur by chance under the null hypothesis. Alternatively, *simulation-based inference methods* model chance outcomes using simulations, which employ physical chance devices (e.g., coins, dice, spinners) or a computer to mimic a random process. Both traditional and simulation-based inference require statistical reasoning (Ben-Zvi & Garfield, 2004) to connect empirical outcomes and hypothesized models, but they differ in the types of models used.

Because traditional and simulation-based inference differ in their use of models, the two approaches may be associated with differences in the way students connect concepts and interpret statistical information. To explore this possibility, a qualitative dissertation study was conducted in the context of an AP Statistics course that uses both traditional and simulation-based inference methods in instruction. Presenting results from that study, this article characterizes how students use inferential models, representations, and tools as they work together to carry out a statistical inference task.

Theoretical Perspective

This exploration of how students use traditional and simulation-based inference methods will be informed by a *models and modeling perspective*, which conceptualizes learning as model-building, an iterative process in which students invent, extend, and revise constructs (Lesh & Doerr, 2003b). In this study, the term *model* will be defined broadly:

> Models are conceptual systems (consisting of elements, relations, operations, and rules governing interactions) that are expressed using external notation systems, and that are used to construct, describe, or explain the behaviors of other system(s)—perhaps so that the other system can be manipulated or predicted intelligently. (Lesh & Doerr, 2003, p. 10)

This definition suggests that models have two parts: conceptual systems in the minds of learners and various representational media (Johnson & Lesh, 2003). Thus, models and representations are intimately related, as "the meaning of a model, or conceptual system, tends to be distributed across a variety of interacting representation systems" (Lesh & Doerr, 2000, p. 239). In this article, the term *representation* will be reserved for external systems, such as graphs, equations, and concrete models. In addition to external representations, this study will consider other tools that support statistical inference, including calculators, computer programs, and problem-solving heuristics.

Central to the models and modeling perspective is the assumption that reality is accessed through models and representations that emphasize different aspects of the underlying system (Lesh & Doerr, 2003b). The discipline of statistics consists of a distinctive set of models and representations of real-world phenomena (Wild & Pfannkuch, 1999), so attention to models and modeling is a powerful lens in this content domain. Lehrer and Schauble (2007, p. 157) argue that "representational change both reflects and instigates new ways of thinking about the data;" thus, it is critical to understand how students interact with the various models and representations used in statistics instruction.

This study conceptualizes traditional and simulation-based inference as two models and corresponding representational systems used to express the logic of inference, and a models and modeling perspective supports the article's central research question: How do students use the models, representations, and tools of traditional and simulation-based inference as they reason about a statistical inference task?

Two Approaches to Statistical Inference

This section introduces the two approaches to statistical inference explored in this study First, a brief literature outlines the proposed advantages and empirical studies that have inspired the use of simulation-based inference in introductory statistics classes. Next, an example is used to contrast the modeling process for the two approaches, highlighting differences in the tools and representational media employed. Finally, this section provides details about the inferential instruction experienced by the participants in this study.

Literature Review

Before modern computing power allowed for rapid simulations, introductory statistics courses necessarily relied on traditional methods like z-tests and t-tests to introduce the core logic of inference (Cobb, 2007). However, at the recommendation of prominent statistics educators, most notably George Cobb (2007), simulation-based inference methods have begun to replace or complement traditional inference methods in a number of introductory courses, including statistics courses at the high school level

(Rossman & Chance, 2014). Statistics educators have suggested that these methods require less prerequisite knowledge, generalize easily to a large number of settings, incorporate modern computing power in a meaningful way, and support conceptual understanding of inference (Chance & Rossman, 2006; Cobb, 2007; Holcomb, Chance, Rossman, Tietjen, et al., 2010).

Evaluations of curricula that use simulation-based inference as the primary means of conducting inference find student performance on multiple-choice CAOS assessment items (delMas et al., 2007) is similar for students who study simulation-based and traditional curricula (Chance & McGaughey, 2014; Garfield et al., 2012; Tintle et al., 2011). However, simulation-based curricula are linked to modest gains on certain topics including modeling and simulation (Garfield et al., 2012), study design and tests of significance (Tintle et al., 2011), and understanding tests of significance as a test of whether observed results plausibly occurred "by chance alone" (Chance & McGaughey, 2014).

Contrasting the Modeling Process across Approaches

The participants in this study were enrolled in an AP Statistics course that employed both traditional and simulation-based inference methods in instruction; thus, these students used both theoretical and empirical models for the outcomes that would occur by chance under the null hypothesis. This section contrasts the modeling process for the two approaches, highlighting differences in the tools and representational media employed. Further, this section provides specific information about the instructional experiences of the participants in this study.

Consider the following example, which necessitates an inference method to determine the statistical significance of experimental results:

> [Earlier in the year] we learned how characteristics of an interviewer can lead to response bias. Two AP Statistics students decided to investigate this issue. They speculated that students would be more likely to identify as feminists if asked by a female interviewer. A sample of 60 male high school students were asked, “Are you a feminist?” Half were randomly assigned to a male interviewer and half were randomly assigned to a female interviewer. Of the 30 asked by a male interviewer, 11 responded, “Yes.” Of the 30 asked by a female interviewer, 15 responded, “Yes.”

Does this result provide convincing evidence of response bias or could the result have occurred merely by chance? There are two approaches to answer this question: a traditional approach and a simulation-based approach.

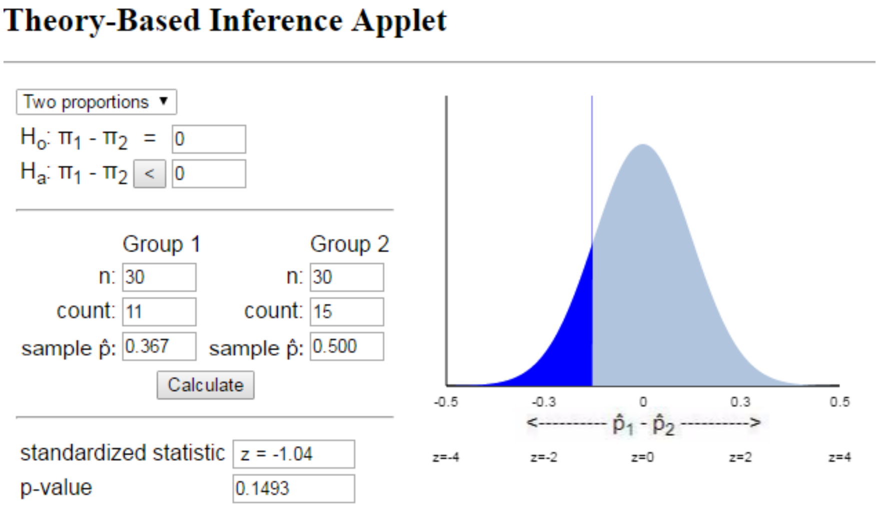
Theoretical Models for Traditional Inference

Suppose we choose a traditional approach to inference. A theoretical probability distribution would be used to model the distribution of outcomes that would occur if the gender of the interviewer had no impact and differences in proportions were due to random chance alone. The sample size is sufficiently large, according to commonly used guidelines, so the Normal distribution can be used as a model. Thus, we can use a *z* test to determine the statistical significance of the experimental results.

In the original experiment, fewer subjects claimed to be feminists when asked by a male interviewer. The difference of proportions, $\hat{p}_M - \hat{p}_F = \frac{11}{30} - \frac{15}{30} = -0.133$, corresponds to a standardized *z*-statistic of $z = -1.04$. When conditions are met, the *z*-statistic has a standard Normal distribution, which can be used to calculate the probability of obtaining a test statistic of $z = -1.04$ or less assuming the null hypothesis is true. This test would result in a p-value of 0.1493. Thus results like those observed in the experiment would be fairly likely to occur by chance even if the gender of the interviewer had no effect on the responses. We cannot reject the hypothesized model, and these results do not provide convincing evidence of response bias.

Alternatively students may consider a chi-square test, which uses a chi-square distribution to model the outcomes that would occur by chance. A large chi-square statistic indicates large differences between the observed counts and the counts that would be expected if there were no response bias. However, a chi-square statistic does not account for the *direction* of the response bias; in other words, a chi-square test is equivalent to a two-sided z-test, resulting in a p-value of 0.2974.

As shown in Figure 3-1, continuous probability models like the Normal, *t*, and χ^2 distributions are often represented as smooth density curves, and the p-value is often represented as an area under the curve by locating the test statistic and shading values that are more extreme in the direction hypothesized by the researchers. [1] Other inscriptions – such as, formal specification of the null and alternative hypotheses – are also common.



Theory-Based Inference Applet

Two proportions ▼

H_o: $\pi_1 - \pi_2$ = 0

H_a: $\pi_1 - \pi_2$ < 0

	Group 1	Group 2
n:	30	30
count:	11	15
sample p̂:	0.367	0.500

Calculate

standardized statistic: z = -1.04

p-value: 0.1493

Figure 3-2. Theory-based inference applet.

[1] The applet used to calculate the p-value is available at http://www.rossmanchance.com/applets/TBIA.html

The AP Statistics curriculum prescribes traditional inference methods, and the course includes nine tests of significance (College Board, 2010). In total, statistical inference – which includes tests of significance and confidence interval estimation – constitutes 30-40% of an AP Statistics course (College Board, 2010). Because traditional inference requires substantial prerequisite knowledge, this approach is typically taught in the final third of the course (Malone et al., 2010); the course textbook, *The Practice of Statistics* (Starnes et al., 2013), adheres to this pattern.

The Practice of Statistics (Starnes et al., 2013) uses the same four-step process for all significance tests:

- State: What *hypotheses* do you want to test, and at what significance level? Define any *parameters* you use.
- Plan: Choose the appropriate inference *method*. Check *conditions*.
- Do: If the conditions are met, perform calculations.
    - Compute the test statistic.
    - Find the ***P*-value**.
- Conclude: *Interpret* the result of your test in the context of the problem. (Starnes et al., 2013, p. 552, emphasis in original)

Over a period of several months, students applied the State-Plan-Do-Conclude process to many traditional inference tasks in class, on assignments, and on tests. Following this framework, hypotheses were stated using formal inscriptions. The appropriateness of a particular test was determined by checking a list of conditions. Calculations were introduced through formulas, but later carried out largely by dedicated inference functions in the TI-84 Plus calculator. Conclusions based on the p-value were always interpreted in context.

Empirical Models for Simulation-Based Inference

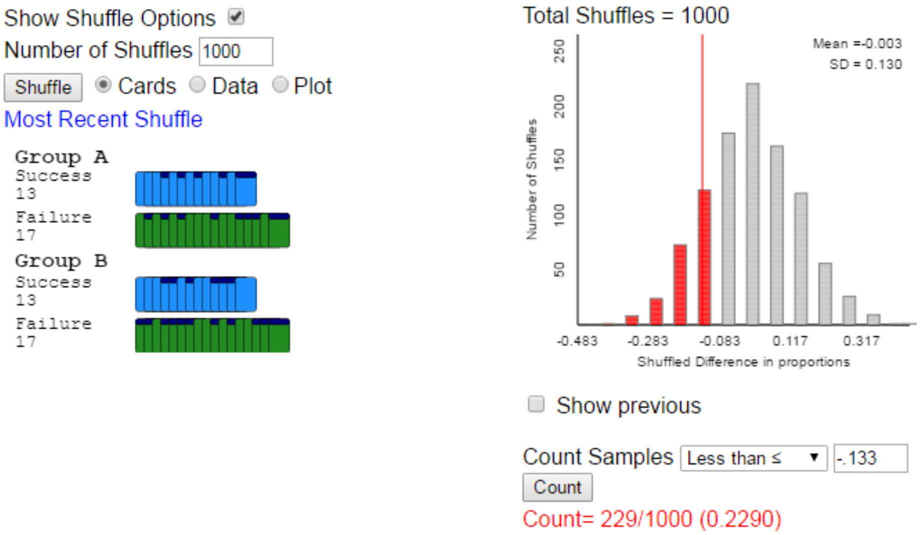
Alternatively, suppose we chose a simulation-based approach to inference. Using a physical chance device, we model the outcomes of the study: for example, let 26 blue cards represent all subjects who said yes and let 34 green cards represent all subjects who said no. We shuffle the cards from both groups together, assuming that the gender of the interviewer has no impact on the response. To model the difference in proportions that would occur by random assignment alone, the cards are shuffled and dealt into two groups. For each trial, we record the difference in the proportions of blue cards in each group. We can use an applet[2] to simulate data for many trials, recording the difference of proportions each time. The results of this simulation are shown in Figure 3-2.

We evaluate the results of the experiment by comparing the outcome of the original study to the distribution of outcomes produced by the model. Out of 1000 simulated trials, there were 229 where the difference of proportions was -0.133 or less, an estimated p-value of 0.229.[3] That is, if we assume the gender of the interviewer had no effect on subjects' responses, a difference as or more extreme than the difference in the original study is fairly likely to occur.

Because simulated sampling distributions are defined in terms of repeating the randomization process many times, they are often represented graphically using dotplots or histograms instead of idealized curves. Further, the vocabulary, notation,

[2] Available at http://www.rossmanchance.com/applets/ChisqShuffle.htm

[3] In this example, the simulation-based approach – which models chance outcomes using a discrete distribution – leads to a p-value that is appreciably different than the one obtained through the traditional approach – which models chance outcomes using a continuous distribution.



Show Shuffle Options ☑

Number of Shuffles 1000

Shuffle ● Cards ○ Data ○ Plot

Most Recent Shuffle

Group A
Success
13
Failure
17
Group B
Success
13
Failure
17

Total Shuffles = 1000

Mean =-0.003
SD = 0.130

☐ Show previous

Count Samples Less than ≤ -.133

Count

Count= 229/1000 (0.2290)

Figure 3-2. Simulation-based inference applet.

and inscriptions may differ from traditional inference. For example, an instructor using simulation-based inference may refer to the “just by chance” explanation instead of formally defining the null hypothesis, though these details certainly vary across instructors. Pfannkuch (2005, p. 280) warns that more informal language may still be misunderstood by students; for example, “The term ‘chance’ should not be lightly overlooked in teaching, as students may understand the term in dice problems but may not for real problems where causes are known.”

In addition to traditional inference methods, the AP Statistics course under study regularly incorporated simulation-based inference methods. Because these methods seem to require less prerequisite knowledge (Cobb, 2007), simulation-based inference can be introduced much earlier in the year. Beginning on the first day of the school year, this class included fourteen in-class simulation-based inference activities, including multiple opportunities for groups of students to design their own simulations. In the second semester, simulation-based inference activities were included as part of every chapter, so students used simulations to model a variety of study designs. The use of simulation-based inference to complement traditional instruction is supported by the course textbook (Starnes et al., 2013), and additional activities were drawn from other sources. Detailed information about these activities is provided in Appendix E.

To help students recognize the unified modeling process of simulation-based inference, the teacher adopted the 3S Strategy used in the *Introduction to Statistical Investigations* curriculum (Tintle, et al., 2013), applying the same approach to each inference task:

> 1. Statistic: Compute the statistic from the observed sample data.
> 2. Simulate: Identify a "by chance alone" explanation for the data. Repeatedly simulate values of the statistic that could have happened when the chance model is true.
> 3. Strength of Evidence: Consider whether the value of the observed statistic from the research study is unlikely to occur if the chance model is true. If we decide the observed statistic is unlikely to occur by chance alone, then we can conclude that the observed data provide strong evidence against the plausibility of the chance model… (Tintle, et al., 2013)

This three-step process was applied to all simulation-based tests of significance throughout the year.

Interpreting P-values

Whether calculated theoretically or empirically, p-values have the same interpretation. Holcomb, Chance, Rossman, and Cobb (2010) identify four components of a correct p-value interpretation: probability of observed data, tail probability, based on randomness, and under null hypothesis. That is, students should be able to indicate that the p-value is the probability of observing results as or more extreme than those observed, identifying what "results as or more extreme" means in the context of the problem at hand. Students should also be able to identify the source of randomness, distinguishing between random sampling and random assignment, and they should understand that the p-value is calculated based on the assumption that the null hypothesis is true (Holcomb, Chance, Rossman, & Cobb, 2010).

Methods

Because of its emphasis on inductive data analysis for generation of theory (Charmaz, 2014) and its compatibility with various epistemological paradigms (Taber, 2000), a modified grounded theory methodological approach was used in this study. This choice had several important implications for data collection and analysis (Charmaz, 2014; Corbin & Strauss, 1990; Creswell, 2013). First, data analysis in grounded theory is always inductive; thus, systematic coding of data in this study relied on constant comparison among data, codes, and categories rather than *a priori* theory. Second, data collection and analysis in grounded theory are carried out simultaneously, and the relationship between the two is expected to be bidirectional. Thus, ongoing data collection was informed by the analysis of previously collected data in order to saturate emerging categories. Finally, grounded theory employs memoing as a means of analysis and theory development. In this study, memos provided a space to develop ideas, plan subsequent data collection, and critically reflect on the research process (Charmaz, 2014).

Participants

This study was situated in the context of an AP Statistics taught by the author at a public school in the southeastern United States. The school is associated with a large research university and serves students in grades K-12. The admissions policy of the school aims to create a student body that mirrors the demographics of the state for five admission categories: gender, race/ethnic origin, family income, exceptional student status, and academic achievement level.

Eleven students from grades 10-12 were enrolled in AP Statistics in the year the study was conducted[4], and all eleven students assented to participate in the study with their parents' consent. A demographic description of the study participants is given in the table below.

Table 3-1. Demographic description of study participants.

Sex		Race		Grade level	
Male	3	White	7	10th	2
Female	8	Black	0	11th	3
		Asian	1	12th	6
		Hispanic	3		

As described above, these participants had been exposed to a curriculum that included both traditional and simulation-based inference methods. However, because the AP Statistics course description emphasizes traditional inference, students had considerably more experience with theory-based methods by the end of the school year. The impact of students' in-class experiences on the analytical generalizability of this study will be discussed later in the paper.

Data Collection

Group interviews

This article focuses primarily on data collected through task-based interviews (Maher & Sigley, 2014). Ten students enrolled in the course were interviewed in pairs. (One student was unable to participate because of absence.) The interviews, which

[4] Additionally, seven students participated in a pilot study two earlier. In this article, no results from the pilot study are presented.

lasted between 25 and 45 minutes, were conducted during the school day in the weeks immediately following the AP Statistics exam. The interviews were audio-recorded and transcribed. Additionally, the students' use of an applet for simulation-based inference was recorded using a screen-capturing tool.

A two-person protocol was chosen for this study for several reasons. First, the interview data were intended to approximate how students reason about inference in a natural classroom setting, and in class, students typically worked on inference tasks in pairs. In particular, classmates who had worked together on similar tasks in class were paired together to help students feel comfortable in the interview environment. Additionally, Schoenfeld (1985) suggests that interviews with multiple students produce rich data for investigating students' problem-solving processes. Multi-person protocols ease the pressure to "produce something mathematical for the researcher," thus eliciting more natural responses (Schoenfeld, 1985, p. 178), and discussions among students makes the reasoning behind their decisions more visible (Schoenfeld, 1985).

In the interview task, students were asked to draw conclusions from the response bias experiment described in the previous section. After reading about the study, students were asked to work together to decide whether the results provide convincing evidence of response bias. At the beginning of the interview students were made aware of all available tools – a formula sheet, a graphing calculator with statistical functions, an extensive collection of chance devices (coins, dice, cards, etc.), and a computer applet. However, no instructions were given about whether students should use a traditional or simulation-based approach. After drawing a conclusion, students were asked to consider the alternative approach; e.g. students who initially carried out a traditional test were asked to carry out a simulation-based test.

Other than the presence of recording devices, this activity was intentionally similar to others that students had completed in class. For example, students were asked to design a physical simulation before using the applet on the computer, as they often did in class. In general, the interviewer intervened in the students' problem-solving process as little as possible, but when students struggled to move forward, the interviewer sometimes prompted them with questions such as, "What are you trying to represent?" or "Would it help you to see the applet?" Again, these were intended to reflect the kinds of prompts students would hear if they asked for help from the classroom teacher. Paper was provided to facilitate communication between partners, but students were not required to produce written work.

Other data sources

Schoenfeld (1985, p. 174) reminds us, "Any framework for gathering and analyzing verbal data will illuminate certain aspects of cognitive processes and obscure others: There are trade-offs between structured and unstructured interviews, single or group protocols, etc." Because, the study described in this article was conducted as part of us a larger dissertation study, it was possible to triangulate the verbal data collected through group interviews with other data sources. These included student responses to targeted formative assessments, exam items, and survey items; daily field notes and journal entries written by the teacher-researcher; and transcripts of individual interviews.

Data Analysis

Data analysis consisted of a process of systematic coding in multiple phases, according to the guidelines for grounded theory presented by Charmaz (2014); these guidelines provided a systematic yet flexible way to study the emerging data. In the initial coding phase, each segment of data was assigned a concrete and descriptive code intended to reflect the students' actions. After comparing these initial codes to the data and looking for patterns in the codes across interviews, the researcher inductively constructed a set of focused codes to “sift, sort, synthesize, and analyze” the large amounts of data (Charmaz, 2014, p. 138). The focused codes reflect the “sensitizing concepts” of the models and modeling perspective, as well as the researcher's familiarity with descriptions of inference by other statistics educators. However, these concepts were not accepted into the analysis until they could be substantiated in the data (Charmaz, 2014; Corbin & Strauss, 1990).

Although students were not required to interpret the p-value as a probability, review of the initial codes revealed that the components of p-value interpretation proposed by Holcomb, et al. (2010) arose frequently in their discussions. To facilitate analysis, these components were included as focused codes: probability of observed data, tail probability, based on randomness, and under null hypothesis.

The focused codes were iteratively applied to the data and revised, until every incident of student problem-solving could be coded. Table 3-2 compares the focused codes applied at the first iteration to those applied at the final iteration.

One code, *summary of observed data*, was removed completely; the researcher had anticipated that students would summarize the results of the study, perhaps by interpreting the difference in proportions for the two groups, but that expectation was not substantiated in the data. Two codes were added: *discussion of study design* and *use of tools*. As described in the Results section, these codes, which the researcher initially imagined were separate from the modeling process, had considerable explanatory power.

Table 3-2. Development of focused codes.

First Iteration	Final iteration
---	Discussion of study design
Summary of observed data	---
Definition of hypotheses	Specification of claims to be tested
Choice of model for chance outcomes	Choice of model for chance outcomes
Evaluation of model fit	Evaluation of model fit
---	Comparison of results to model
Conclusion	Decision about claims
Use of graphical representation	Use of graphical representation
---	Use of tools
Evaluating strength of evidence	---
Probability of observed data	Probability of observed data
Tail probability	Tail probability
Based on randomness	Based on randomness
Under null hypothesis	Under null hypothesis

Another code was changed substantially. When components of the p-value interpretation – *probability of observed data*, *tail probability*, *based on randomness*, and *under the null hypothesis* – arose in the initial codes, the researcher hypothesized a cover term – *evaluating strength of evidence* – that would include these components. However, focused coding revealed that these codes did not arise from formal evaluations of evidence at the point of drawing a conclusion. Rather, they arose at various points during the modeling process. For example, the following exchange between Ryan and Eva was coded *under the null hypothesis*; this concept arose as they evaluated competing physical models for the simulation, not at the conclusion phase.

Ryan: That would be testing like – assuming that everybody – because that would be assuming like there isn't response bias and that what they say is true, and then you'd be doing…

Eva: Aren't we supposed to do that though?

Ryan: I'm going to take out four pinks.

Eva: Wait, I thought we were supposed to test assuming…

Ryan: Yeah, and that’s what it would be doing if you took all the yeses and the nos, because if you assume that there’s no response bias, then in the sample 11 people were actually saying yes…

Eva: Oh, I see what you’re saying.

Even at the point of estimating a p-value, which the 3S strategy conceptualizes as measuring strength of evidence, students rarely used the language of strength/weakness. Consider Alicia’s reasoning as she calculated a p-value using the applet.

Alicia: Ok, so our difference – our observed difference from this sample is 0.133. So like where that would fall on that graph is kind of in the middle, between 0.117 and 0.317.

Catherine: Over here somewhere?

Alicia: Yea, in that third bar maybe? We can – so let’s draw that just so I can see it. So we have our little graph. So our observed difference is 0.133 here, and… Wait, no we use that number (referring to the most recent shuffle). … No, we don’t care about that number. We care about this number (referring to the observed difference in proportions). So we look where 0.133 falls on the graph, and that’s our p-value.

Notice that she never described what she was doing in terms of the strength of evidence. Thus, the label *comparison of results to model* was judged to be a more faithful rendering of students’ actions.

Similarly, other codes were revised to better reflect the empirical data. For example, *definition of hypotheses* connotes a formal definition, perhaps writing both the null and hypotheses using symbols. In contrast, some students specified the claim to be tested more informally.

Devon: So what would response bias look like? So we’re testing response bias against no response bias. So I guess we could like model a distribution of if there was no response bias.

To better reflect these data, the code was changed to *specification of claims to be tested.* The results section includes examples of each code.

Results

The data suggest that there are qualitative differences in how these students use the models, representations, and tools of the two inference methods. A few broad differences provide context for the more subtle differences presented later in this section. First, all five pairs chose to complete the task using traditional inference first. Compared to the simulation-based approach, students appeared to apply traditional inference methods with more confidence and less effort, and all pairs were able to reach a reasonable conclusion without intervention by the interviewer. In contrast, completing the task using simulation-based inference required considerable discussion among students and some prompting from the interviewer. Besides having more experience with the traditional approach, some students volunteered reasons that they find it easier than simulation-based inference, calling traditional inference "straight to it" and "structured."

The second broad difference is related to the first. For all five pairs, the part of the interview devoted to simulation-based inference was much longer than the part devoted to traditional inference. In addition to the reasons given above, recall that students were asked to design a physical simulation before using the applet to carry out simulation-based inference. There is no analogous design step in traditional inference; instead carrying out traditional inference requires students to choose a theoretical probability distribution and evaluate its fit for the given data.

This section presents examples of each focused code, comparing and contrasting incidents across approaches. For each approach and each pair of students

(P1-P5), Table 3-3 outlines the conceptual content of the discussion; a check mark indicates the code was applied to that pair's transcript. One salient trend revealed in the table is the absence of p-value interpretation codes among students using traditional inference methods; none of these five pairs of students explicitly mentioned the probability of observed data, tail probability, the source of randomness, or the assumption that the null hypothesis is true while conducting and interpreting a traditional significance test. Implications and limitations of this finding are discussed later in the paper.

Table 3-3. Incidence of focused codes in group interviews.

	Traditional					Simulation				
	P1	P2	P3	P4	P5	P1	P2	P3	P4	P5
Discussion of study design		✓	✓						✓	
Specification of claims	✓	✓	✓	✓	✓			✓	✓	✓
Choice of model	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Evaluation of model fit		✓	✓			✓	✓		✓	✓
Comparison of results			✓		✓	✓	✓	✓	✓	✓
Decision about claims	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Use of graphical representation			✓			✓	✓	✓	✓	✓
Use of tools	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Probability of observed data						✓	✓	✓	✓	✓
Tail probability						✓		✓	✓	✓
Based on randomness							✓		✓	
Under null hypothesis						✓		✓	✓	✓

Students' Discussion of Study Design

Three of the five pairs of students mentioned study design in their discussion of this task – two while using traditional inference and one while using simulation-based inference. The pairs using traditional inference were both using the State-Plan-Do-Conclude framework. As part of the Plan stage, students routinely check a list of conditions for inference, which includes the assumption that the data were randomized.

Grace: Ok, so (reading), this was – it was a sample.

Zoe: But there's Random.

Grace: Yea, but it was an experiment, because they were randomly assigned. Because like the treatments are the male and female interviewers.

Only one pair of students mentioned study design while using simulation-based inference. William and Libby had just decided to use a model that re-randomized the original data, when William noticed that their model would always result in more subjects saying no than saying yes, because there were more nos than yeses in the original sample. The following quote raises a question related to the scope of inference in direct connection to the construction of their model.

William: Oh, yea, but that doesn't seem like a very efficient one, because what if the population is 50-50? I mean, this is just one sample. That's basically – you're taking the data from this sample and just shuffling it around.

This led to a discussion about what generalizations are appropriate based on various study designs.

Specification of Claims

As they carried out traditional inference, four of the five pairs explicitly specified the claims to be tested. Of those, three used a formal inscription like the following:

$H_0: p_2 - p_1 = 0$, $H_A: p_2 - p_1 > 0$. These inscriptions are part of the State phase in the State-Plan-Do-Conclude framework.

Ryan: So if we had a hypothesis for this, it would be that the amount who said yes – we could use that as a proportion out of 30, so p_1 the number who said yes when interviewer male is equal to p_2 the number who said yes when interviewer female.

Eva: Nice.

Ryan: The alternative, you would expect the male interviewer to get more yeses.

Eva: Yeah, so p_1 is less than p_2.

In contrast, only one pair explicitly stated their hypotheses while using a simulation-based approach. However, this may be a consequence of the groups choosing to carry out traditional inference first. Data from the larger dissertation study suggest that these students readily accepted traditional and simulation-based approaches test the same hypotheses (more details in Chapter 5).

Choice of Model and Evaluation of Model Fit

Comparing incidences of this code across the two approaches revealed substantive differences. Students using traditional inference methods chose models implicitly, by choosing a particular significance test. The justification for the decision typically mentioned the type of data (categorical), the summary statistic (sample proportions), and/or the number of groups.

Grace: So these are proportions because it's like 11 out of 30 and then like 15 out of 30.

Zoe: So a z-test.

Grace: Yea.

Zoe: Two proportion z-test.

Two pairs of students evaluated the fit of the theoretical probability model by checking conditions. Both did this as part of the State-Plan-Do-Conclude framework, and neither verbalized the connection to the shape of the sampling distribution. In the incident below, Alicia and Tianna try to remember how to check the Normal condition, and their discussion of these details continued beyond what is reproduced here.

Alicia: Plan. Ok, if it's Normal.

Tianna: n is greater than 10.

Alicia: Proportion is if it's if np is greater than 10, right?

Tianna: Right, yea, yea, yea. And then the n (p-1) or 1 minus…

Alicia: No. Yea, n (1-p)

Tianna: Yea.

Alicia: Yea, that makes sense. Is it greater than or equal to 10? I think so. I think it's greater than or equal to. It's not going to be 10, chances [are].

Other data sources from the larger study confirm that difficulties with the technical details of checking conditions for inference were common throughout the course. Of the groups who did not check conditions, none expressed doubt about the validity of their eventual p-value based on concerns about model fit.

The choice of model and evaluation of model fit were considerably different as students used a simulation-based approach to inference, which required them to design a physical simulation before using the applet. First, these two codes were likely to appear in a cyclical pattern as students proposed and evaluated models, rejected those that did not fit well, and proposed new models. Second this phase of the inference process often elicited conversations about the need to model the assumption of the null hypothesis. The incident below is Alicia's justification to her partner for a model using 26 red cards to represent the subjects that said yes and 34 green cards to represent the subjects who said no.

Alicia: Ok. So if you think about it we're claiming that the gender of the interviewer doesn't affect whether they say yes or no. So we're claiming… Actually they're claiming that gender does affect it. But we're just trying to test the null. Ok, so we're testing the null, and the null is that gender doesn't affect the response. So we're saying out of – according to the sample, out of 60 people, 26 said yes and 34 said no completely independent from the gender.

Students also sometimes mentioned that the model chosen was based on randomness, though this was more difficult to code. Several pairs noted that their model was a representation of a "just by chance" explanation – a phrase used often in classroom instruction. However, it was not always clear if the phrase referred to the null hypothesis or the source of randomness; it is possible some students understand "just by chance" as shorthand for both of these components. Some students shuffled cards and dealt them into two groups, but did not explicitly link the dealing to random assignment. Implicit references to random assignment were marked in the transcript but were not coded as "based on randomness." Other pairs made the meaning of "just by chance" more explicit, as in the incident below:

Libby: We need 60 cards. So we should shuffle these together – shuffle these together so they're in a random order. And then do male interviewer, female interviewer, male interviewer, female interviewer (mimes dealing cards).

William: Or we could just do male interviewer, female interviewer whatever it is (mimes counting out 30 shuffled cards for the pile). Let's see if we can get these results just by chance.

Comparison of Results to Model

While using traditional methods, only two pairs of students explicitly compared the empirical results of the experiment to the hypothesized model. In contrast, the applet for carrying out simulation-based inference essentially requires it. Most instances of the “probability of observed data” and “tail probability” codes accompanied the comparison of the original experimental results to the distribution created by simulation.

Devon and Isabella described both traditional and simulation-based methods in terms of comparing their results to a reference distribution.

Devon: df = 1, chi-square = 1.0, p-value =0.29. And then you fail to reject the null hypothesis, because it might as well fall into this distribution.

Unlike others who struggled with the applet, this pair immediately drew conclusions based on comparison of the original results to the simulated distribution.

Decision about Claims

After calculating a p-value, students made a decision about the claims being tested by comparing the p-value to a common significance level; in this example, they failed to reject, because the p-value was relatively large.

Ryan: Okay, so our chi-squared value is 1.08. P value is 0.297, not negative. Degrees of freedom equals one.

Eva: Yea.

Ryan: So then if we set our hypothesis at alpha equals 0.05 then…

Eva: Looks like it’s greater than that.

Ryan: So then we fail to reject.

Eva: We sure did.

Ryan: There is insufficient evidence of response bias in this study. There you go.

These final conclusions were nearly identical across approaches. Once students recognized the probability calculated as a p-value, they were quick to apply the familiar rule. In several cases, students initially misapplied the rule – they intended to reject the null, because the p-value was large – but they were always corrected by their partners.

As discussed above, many students discussed the components of p-value while evaluating model fit and comparing their results to the model, but they did not directly connect these components to the decision about claims. Only Isabella linked components of the p-value interpretation to a logical conclusion, but she shared this reasoning during the modeling phase before the p-value had been estimated.

Isabella: So this is the null. So if this is like the – how would you say it – the distribution of the null hypothesis, then this wouldn't matter, the gender of the interviewer. And so then if this – if the p-value out of this distribution is small, then that would mean that the gender of the interviewer does affect the answer that these male high school students would give.

Use of Graphical Representations

While carrying out traditional inference, only one pair of students made use of a graphical representation of the sampling distribution. Alicia drew a density curve to explain to Tianna how she remembers whether to reject or fail to reject when the p-value is small. As shown in the explanation below, Alicia's memory aid makes an oblique comparison of the results to the hypothesized model, but it is not based on a conceptual interpretation of the p-value.

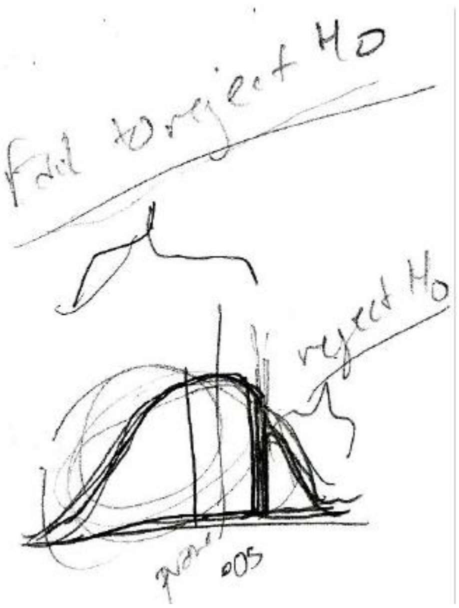


Figure 3-3. Student work.

Alicia: Oh, I just said like on a normal graph, you put wherever – if you're trying to decide whether to fail to reject or just to reject the null, you put the alpha level on the graph, and then you write wherever the p-value is. So if the p-value falls in the bigger part, you fail to reject. And I just thought of it like fail to reject is a longer sentence, like that's longer to write. Or if it falls in the little part, then it's just reject.

To earn full credit on the AP Statistics exam, it is not necessary to create a visual representation of the p-value. These data suggest that students may not spontaneously use graphical displays to aid in their reasoning when such representations are not required.

The applet used to carry out simulation-based inference automatically creates a dotplot of the simulated sampling distribution. The mere appearance of the graph on the screen does not constitute use of graphical representation, but verbal data suggests that students did incorporate the graphical display in their reasoning.

Grace: So it would be like 4 over 30 would be the difference [in the sample proportions], and so we found that – oh, and that's right there actually. And so that's that. So we can put that here, because that's the number we're testing for to see if – see how likely it is just by chance.

As shown in Table 3-3, all pairs of students explicitly referenced the graphical representation as they reasoned about the simulation-based approach.

Use of Tools

Although traditional inference was initially introduced in class using formulas for the test statistic, by the end of the year, all five pairs of students had transitioned to the use of calculators exclusively. Though some individual students had questions about how to use the calculator, all pairs were able to calculate a p-value without any support from the interviewer. The TI-84 Plus calculators include wizards that prompt students to enter all necessary information for a given statistical test. Beyond reducing syntax errors, the researcher observed that these wizards were sometimes used by students to help them choose a test. This use of the calculator to support the choice of test is confirmed by data from the larger study. For example, when asked a question about calculator use in his individual interview, Devon offered the following explanation:

Devon: I'd just keep trying a bunch of tests until I found, oh this one. And then it'll be like between two tests, and then I'll decide.

Catherine: How do you know from the calculator? How does that help you decide?

Devon: It has certain options, so I know it's not a t-test. I can narrow it down. Because sometimes there's like no way I can get a df or – I don't know. There's just ways to narrow it down on the calculator. Because sometimes you won't have the information necessary.

Further, the calculator wizard may prompt students to adjust their hypotheses in some cases. Originally, Alicia and Tianna had stated their hypotheses as $H_0\colon p = 0.5$ and $H_0\colon p > 0.5$, where the proportion referred to the students who say they're feminists if asked by a female interviewer. However, using the calculator as a guide, they chose a two-proportion test, and while entering their data into the calculator, Alicia recognized the need to modify the null hypothesis.

Alicia: (speaking out loud as she uses the calculator) Then p_1, we go to our alternative, p_1 is greater than… oh, that's what we're missing! p_1… So in our null we're saying that p_1 equals p_2. So the proportion of people who say they are feminists to a male interviewer will be the same as the proportion of people who say they're feminists to a female interviewer. So that's our null.

In other cases, the calculator allowed students to calculate a p-value without confronting uncertainty about the hypotheses being tested. In particular, two pairs of students who chose to use a chi-square test were able to calculate a p-value and come to a conclusion about response bias in the study, without correctly conceptualizing response bias in terms of proportions. Specifically, they did not recognize that a proportion of self-identifying feminists different than 0.5 is not, in itself, evidence of response bias. An illustration of this is given in the next section.

As discussed previously, the applet often prompted students to discuss the p-value as the probability of observed data or as a tail probability as they used the tool to estimate a p-value. However, these students were not nearly as proficient with the applet as they were with the calculator. In addition to technical questions about the interface, some uses of the applet suggested substantively different conceptions. For example, the group interviews included incidents in which students interpreted a single simulated sample, treated simulated data as real data, and combined traditional and simulation-based approaches. These issues are described in detail in Chapters 4 and 5.

Discussion

Which Approach is “Easier”?

Persuasive arguments have been made for the relative simplicity and intuitiveness of simulation-based inference as compared to traditional inference (e.g., Cobb, 2007; Lock et al., 2014). In particular, simulation-based inference methods do not rely on theoretical models for the sampling distribution – models which Cobb (2007, p. 7) argues are “conceptually difficult, technically complicated, and remote from the logic of inference.” In fact, even in seemingly simple data settings, the statistical theory necessary to evaluate “the fit between model and reality” can be prohibitively complex.

Nevertheless, the participants in this study carried out the traditional inference task more easily than the simulation-based inference task, and several expressed views on the relative ease of the two approaches that conflict with prevailing wisdom in the field. Analysis of the group interviews suggests that students were largely shielded from the technical details of statistical theory as they interacted with the models, tools, and representations of traditional inference. Choosing a statistical test based on basic data features and checking a memorized list of conditions may render the modeling process nearly invisible to students. In contrast, designing a physical simulation makes the work of selecting and evaluating model fit explicit, often prompting consideration of the four components of a p-value interpretation. As students describe their preference for traditional tests, some teachers may hear reasons to include simulation-based inference in their courses.

Isabella: Oh, man. See, I'm not good at simulations. I always go the other way.

Catherine: What makes the simulations harder?

Isabella: I don't know. Just like planning it out, because there's a lot of options to do simulations. It's like the coin flip or the dice roll or the spinner.

Devon: Yea, it's more risky. It involves more thought. The test is just like – it does it for you, so you just put in the data. So if you know when to do the test and you know how to do the test, you don't really have to understand what you're doing.

Use of Tools and Representations

As teachers, it can be tempting to assume that students' problem-solving process is linear. They first conceptualize the problem – consider a statistical question, plan an analysis approach, etc. – and then use tools to carry out their plan. However, these data suggest that the connection between students' reasoning and tools is more complex: Not only do students *use* tools, they *interact* with them. For example, a data analysis tool may serve as a memory aid or as source of cognitive dissonance when the context is not fully understood; a tool may prompt students to discuss certain statistical concepts while enabling them to disregard others. In addition to other factors like availability and ease of use, the impact on student thinking is a nontrivial consideration in choosing tools.

In traditional statistics courses that do not allow graphing calculators, students may be instructed to draw and shade a density curve before reading a p-value from a table. However, when using an inference-capable calculator, students can easily come to the p-value without considering any graphical representations. These interviews suggest that when student-constructed graphical representations are not required, students may not spontaneously produce them. However, when graphical representations are presented through an applet, the same students can incorporate them productively, particularly in consideration of p-values as the probability of observed data and a tail probability. The larger dissertation study included an informal survey question asking students whether they believed experience with simulations had helped them understand z-tests and t-tests. Of the ten students who responded, seven said yes, and the most common reason given was appreciation for the visual representations.

Making Modeling Explicit, Making Thinking Visible

After comparing data and focused codes across interviews and across inference approaches, a salient pattern emerged. Interactions with the models, representations, and tools of simulation-based inference often made modeling explicit, and thus made student thinking visible. The data presented in the previous section highlights how students made their thinking visible to each other through competing models, thus challenging existing statistical conceptions. The visibility of student thinking during interactions with simulation-based inference methods may also have implications for teaching and research. This is consistent with a models and modeling perspective, which posits that "an emphasis on models tends to render student thinking highly visible to teachers" (Lehrer & Schauble, 2006, p. 383) and "model-eliciting activities often function as thought-revealing activities that provide powerful tools for teachers and researchers" (Lesh & Doerr, 2003a, p. 31). The following section illustrates how statistical modeling can be leveraged as a thought-revealing activity.

Statistical Modeling as a Thought-Revealing Activity

This section focuses on the interaction of one pair of students, William and Libby. Their interview was chosen for several reasons. Perhaps most importantly, these students were comfortable working together and disagreed with each other frequently and freely. They proposed and evaluated several competing models, and in doing so, articulated a number of important distinctions for statistical modeling. Further, as measured by a comprehensive course exam administered two weeks before the interview, William and Libby represented average achievement and below average achievement, respectively. The exam was an AP Statistics practice exam provided by College Board and scored according to their detailed rubric; on the AP scale of 1 to 5, William's practice exam received a 3 and Libby's received a 1. These students were engaged as they reasoned about the task and seemed open to revising their statistical conceptions over the course of the interview.

Traditional Inference

Libby and William first approached the task using traditional inference methods, choosing to use chi-square test. Their discussion of the task from beginning to end is reproduced below to allow for comparisons in context. In the transcript below, ellipses stand in for off-topic discussion; no interactions coded for statistical reasoning are omitted.

Libby: Ok, so, here we have our case.

William: I think we should do a chi-square test.

Libby: I disagree. I'm just kidding. I think you're correct. Do you want me to read the numbers to you?

William: No, I think I can do it. (looking at the calculator)

…

Libby: Are we categorizing the groups as like whether they were interviewed by a male or female interviewer or whether or not they say they're a feminist?

William: (working on the calculator) Does that look good?

Libby: Yea.

William: Let's go with… So expected, how many are in each category? Like how many people?

Libby: In total?

William: No, just… Ok, so, 30, 30 would be 15, 15, 15, 15. (reading aloud as he navigates the calculator's statistical functions). I gotcha.

Libby: Wow. So we should probably record all that.

…

Libby: We have our chi-square value, we have our p-value, we have our degrees of freedom. … What do our – What does our chi-square value and what does our p-value mean?

William: P-value means because it's below … let's do 0.05. The p-value 0.29 is greater than the alpha level which is 0.05 so I'd say that we can reject the null. And we should probably come up with a null.

Libby: Our null would be – so we're trying to see if students would be more likely to identify as a feminist if asked by a female interviewer. So the null would be they are not more likely to answer yes. The null would be there is no correlation. … I should say association (looking at interviewer).

Catherine: Mmhmm, that's better.

Libby: So our p-value is greater than 0.05, so we reject the null.

William: Fail to reject the null.

Libby: Oh, yea, we fail to reject then null, and so that means that there could potentially be no association between whether or not they said yes when asked by a female interviewer.

Notice that the chi-square test – with its corresponding use of the chi-square distribution as a model for chance outcomes – was chosen without much discussion or evaluation of model fit. Libby seemed to have questions about how the data were “categorized,” but William proceeded with the calculations for the test without addressing her concerns. Further, notice that William asked about expected counts and went on to enter 15, 15, 15, 15 into the calculator. That is, he expected half of the 30 subjects interviewed by a male interviewer would identify as feminists and half of the 30 subjects interviewed by a female interviewer would identify as feminists if there were no response bias. As discussed previously, this is inconsistent with a null hypothesis of no response bias. No response bias implies that the proportion of subjects who identify as feminist is independent of interviewer gender; it does not imply that the proportion who identify as feminist is 0.5. Ultimately, this error did not affect their p-value, because the TI-84 Plus calculator does not require the user to enter expected counts for a chi-square test of association. Rather the expected counts appropriate for a chi-square test of association were calculated automatically, and the counts William typed in were overwritten. The p-value produced by the calculator corresponds to the appropriate expected counts. Initially, Libby misapplied the rule, rejecting the null hypotheses based on the large p-value, but William corrected her without explanation. Finally, the students arrived at a conclusion and interpreted the results of the test in context.

Simulation-Based Inference

While using traditional inference, William and Libby decided on a model without much discussion and proceeded quickly to a p-value and an interpretation. In contrast, they struggled considerably to choose an appropriate model to simulate outcomes that would occur due to random assignment alone. Several important modeling distinctions arose in their discussions. Data from other group interviews and entries in the teacher-researcher’s journal confirm that each of the three issues described below appeared frequently among this group of students.

Representing outcomes versus representing treatment groups

In one proposed model, Libby and William planned to use an equal number of cards of each color. Libby was concerned, because she noticed that the total number of subjects who say yes was not equal to the total number who say no; she was considering the row totals. However, William believed the model was appropriate, because equal numbers of subjects were assigned to the male and female interviewers; he was considering the column totals.

Table 3-4. Data for group interview task.

Feminist?	Male Interviewer	Female Interviewer	Total
Yes	11	15	26
No	19	15	34
Total	30	30	60

Libby: Oh, wait I should probably use 60 [cards].

William: There's 60 total.

Libby: But but but but it's not equal for like yes and no. It's not the same number of people who say yes…

William: It's the same number of male and female interviewer. 30 were asked by the male, 30 were asked by the female.

This incident calls attention to two similar but distinct random processes: the original random assignment of treatments to groups and the simulated re-randomizing under a "just by chance" assumption.

Data from the larger study confirm that students often struggled to decide whether to represent outcomes or represent treatment groups. In class, this issue was more visible when the sample sizes for the treatment groups were not equal. For instance, suppose the response bias experiment had been randomized by flipping a coin: heads, the subject was asked by male interviewer, and tails, the subject was asked by a female. This could result in unequal samples – say, 28 and 32. If students proposed a model using 28 cards of one color and 32 cards of another color, their confusion about whether to model outcomes or treatment groups would be visible to a teacher or researcher.

Representing random assignment versus random selection

In another proposed model, William and Libby were using a large stack of cards of two different colors (presumably, the same number of cards of each color, given how the cards were packaged). They let one color represent people who said yes and one color represent people who said no, assuming at this point in the interview that the two responses were equally likely under the null hypothesis. When they opened the applet and compared their model to the one shown on the screen, Libby called attention to the distinction between random assignment and random selection.

Libby: (referring to their physical model) Is that what's random? Like your response? [The applet] was testing for, the response doesn't change, but like someone who's a feminist might just randomly be placed with a male interviewer versus a female interviewer.

By dealing from the large stack of cards, William and Libby were randomly selecting subjects with certain responses from a larger population. In contrast, the applet's model, described earlier in this paper, assumed that the responses were fixed and subjects with these responses were randomly assigned to groups.

This distinction is subtle, and in fact, traditional inference methods use models based on random sampling even when the study design is based on random assignment. Cobb (2007, p. 8) famously called this use of the sampling model "fraud": "Do we want students to leave their brains behind and pretend, as we ourselves apparently pretend, that choosing at random from a large normal population is a good model for randomly assigning treatments?"

In this study, Isabella and Devon made visible the same thinking about randomization. They used a spinner divided into portions of 43% and 57% to model subjects' responses, because in the original data $\frac{26}{60} = 43\%$ of the subjects said yes. In their model, like Libby and William's model above, the responses themselves are random, not just the assignment to groups. Data from the larger study also provide examples of students making the opposite error. In some cases, students built models that re-randomized the original sample, even when the original data were not produced through random assignment.

"Just by chance" versus "equally likely"

As they continued to examine the model provided by the applet, William finally addressed the conflict between two different conceptualizations of response bias. Up until this point, William had been conflating the concept of "just by chance" with "equally likely".

Libby: So [the applet model] would have been like if we took – if we said, ok, 26 people in total say yes and we counted out 26 green cards (illustrates by counting out a few cards) and then we counted out 34 red cards.

William: And then assigned them to groups.

Libby: And then randomly assigned them to groups.

William: Oh, no, that's not at all what we were doing [in our physical model].

Libby: Right, so what's the difference in what we were testing?

William: Because that's – your results, you're always going to get more no than yes. Wait, let me see? Yea, you're always going to get more no than yes in that. But in our thing, because it's an equal number of cards you could – we're testing for by chance – that one there's always going to be bias.

Libby: But I think what [that applet] is testing for is that their response is something that doesn't change based on – based on whichever gender. Like that person, no matter who they saw, would answer yes or no.

William's comments reveal his understanding that a "by chance" model should render both responses equally likely; in fact, he believes that a model where the two responses are not equally likely will introduce bias. Libby offers a different interpretation of "by chance" – the assumption that the gender of the interviewer didn't matter.

In other group interviews, students made this conception visible by choosing a coin to model the subjects' responses. Devon and Tianna both proposed using a coin, but in each case, their partners recognized that "just by chance" is distinct from "equally likely" and proposed alternative models. The process of choosing and evaluating a physical simulation model can make students' thinking about this important distinction visible to classmates, a teacher, or a researcher.

Implications

After a year of instruction in an AP Statistics class that employed both traditional and simulation-based inference methods, the participants interacted with the models, tools, and representations of the two approaches in qualitatively different ways. Pairs of students in a group interview setting were able to carry out a traditional significance test quickly and efficiently without intervention from the interviewer. Using tools such as textbook problem-solving frameworks, lists of conditions to evaluate model fit, and graphing calculators with wizards for statistical tests, they avoided the technical complications of traditional inference. However, they also avoided explicit discussions of modeling and of conceptual components of inference in their discussions. This does not imply that the participants were unaware of the theoretical underpinnings of the traditional model; rather their interactions with this inference task – similar to many they completed in preparation for the AP Statistics exam – did not lead to spontaneous discussion of these concepts. In particular, the task did not provide many opportunities for the students to challenge each other’s statistical conceptions or for a teacher-researcher to evaluate student thinking.

In contrast, these students struggled to complete the inference task using simulation-based methods. Although they had completed many similar inference tasks over the course of the school year – a few in very similar data settings – they found the task of designing and interpreting a simulation challenging. As they struggled to find an acceptable model, they often discussed the assumption that the null hypothesis is true and the need to model a “by chance” explanation. As they interacted with the graphical representation provided by the applet, they often discussed the p-value as the probability of observed data and a tail probability. The simulation task required explicit attention to modeling and prompted students to make their thinking visible to each other and to the teacher-researcher.

This study provides a naturalistic description of the use of models, tools, and representations in a single classroom environment; it is not an experimental design. Thus, the results presented do not provide a basis for comparing the effectiveness of traditional and simulation-based inference, since all participants were exposed to both approaches in instruction. However, this study may have implications for teachers as they consider which inference methods to include in their courses, which tools to provide to students, and which inference tasks to pose. Simulation-based methods present a considerable challenge to students, even after a substantial investment of instructional time; however, a simulation approach to inference tasks can prompt students to make modeling explicit and make their thinking visible. When they make their thinking visible to teachers and classmates, opportunities arise for students to construct and revise conceptions of statistical inference.

CHAPTER 4
COMMON ERRORS IN SIMULATION-BASED INFERENCE

A rich understanding of inference is an important outcome of an introductory statistics course. The American Statistical Association's *Guidelines for Assessment and Instruction in Statistics Education: College Report* (ASA, 2005) mentions understanding of statistical inference as a key feature of what it means to be statistically educated. Some statistics educators (e.g., Chance & Rossman, 2006; Cobb, 2007; Lock et al., 2014; Pfannkuch, 2005) believe that simulations have the potential to develop a deeper conceptual understanding of statistical significance and p-values, and today simulation-based inference methods are increasingly common in introductory statistics courses as a complement or substitute for traditional inference (ASA, 2016; Rossman & Chance, 2014).

Adoption of these methods is inspired by exciting proposed advantages and "a generation of adventurous authors" (Cobb, 2007, p. 13) who have published curricula and resources to support implementation of simulation-based inference methods. Further, developers of curricula that employ simulation-based inference as the primary means of teaching inference have published evaluations that suggest students in simulation-based courses compare favorably to students in traditional courses (e.g., Garfield et al., 2012; Tintle et al., 2012, 2011). However, simulation-based inference is not a panacea. This study identifies errors that commonly arise among students who use simulation-based inference methods and characterizes the statistical conceptions underlying those errors. This characterization of student conceptions necessitates a new framework for conceptualizing the logic of inference.

Literature Review

To provide context for the present study, this section presents the proposed advantages and early empirical evaluations that have contributed to the popularity of simulation-based inference in introductory statistics classes. The section concludes with a list of errors that arise in classes that employ these methods; each will be explored more deeply in this study.

Proposed Advantages of Simulation-Based Inference

There are several proposed advantages of using simulation to teach statistical inference. First, the simulation-based approach requires less prerequisite knowledge of probability and no distributional assumptions (Cobb, 2007). Since a simulation-based approach avoids mathematical formulas and theoretical sampling distributions, students may see the connections between data production, model and inference more easily (Cobb, 2007; Lock et al., 2014). The relative simplicity of this approach also allows inference to be introduced early in an introductory course and reinforced in various contexts, whereas traditional inference cannot be introduced without the machinery of theoretical sampling distributions (Holcomb, Chance, Rossman, Tietjen, et al., 2010; Tintle et al., 2011). Second, it is trivial to change the statistic of interest (Holcomb, Chance, Rossman, Tietjen, et al., 2010) and the process easily generalizes to a large number of settings. Third, it incorporates modern computing power in a meaningful way. Not only does it take advantage of the pedagogical uses of technology, as it uses simulations to make abstract concepts more concrete (Chance & Rossman, 2006; Lock et al., 2014), but it modernizes the content of the introductory statistics courses to reflect technological advances (Cobb, 2007; Holcomb, Chance, Rossman, Tietjen, et al., 2010).

Empirical Studies

In addition to philosophical arguments, researchers have begun to empirically evaluate the impact of simulations on students' understanding of inference. Evaluations of curricula that primarily use simulation-based inference find student performance is similar for students who study simulation-based and traditional curricula (Chance & McGaughey, 2014; Garfield et al., 2012; Tintle et al., 2011). However, simulation-based curricula are linked to modest gains on certain topics including modeling and simulation (Garfield et al., 2012), study design and tests of significance (Tintle et al., 2011), and understanding tests of significance as a test of whether observed results plausibly occurred "by chance alone" (Chance & McGaughey, 2014). Though many find these results promising, not all proposed advantages of simulation-based inference are substantiated in these studies; in particular, many students continue to struggle with conceptual interpretations of p-values.

The studies mentioned above feature quantitative analysis of student performance on multiple-choice assessments, specifically the CAOS assessment (delMas et al., 2007). Because the CAOS assessment is not specific to simulation-based inference, it provides a useful metric for comparing the two approaches. However, it not does provide information about errors that commonly arise in courses that employ simulation-based inference. Though large-scale evaluations have not provided theory to explain how novices employ simulation-based inference, developers and users of these curricula have shared brief recommendations regarding errors that commonly arise. These have often been disseminated in the form of conference papers and presentations. Common errors and conceptions that have been reported include the following:

- Difficulty designing or identifying appropriate simulations (Chance & McGaughey, 2014)
- Misidentifying observational units in a simulated sampling distribution (Rossman & Chance, 2014; Saldanha & Thompson, 2002)
- Conflating simulation and replication (Chance & McGaughey, 2014; Hodgson & Burke, 2000; Rossman & Chance, 2014)
- Reasoning that the null hypothesis cannot be rejected because the simulated distribution is centered at the null value (Gould, Davis, Patel, & Esfandiari, 2010)
- Failing to recognize the role of the null hypothesis in the simulation process or the purpose of the simulation (Chance & McGaughey, 2014)

Chance and McGaughey (2014, p. 6) warn, "Don't underestimate the difficulty students may have in the transition from one 50/50 proportion to other scenarios, including the distinction between sampling and assignment." Chapter 3 described the various ways student-designed models can reveal student thinking. This chapter focuses on errors that are common to all simulation-based tests, including tests of a single proportion.

Methods

Context

This study was situated in the context of an AP Statistics course taught by the author at a public school in the southeastern United States. The AP Statistics course description includes four major topics (College Board, 2010): data analysis and exploration (20-30%), study design (10-15%), probability and simulation (20-30%), and statistical inference (30-40%). In addition to the prescribed AP Statistics curriculum, the course regularly incorporated simulation-based inference methods. Beginning on the first day of class, simulation-based inference activities were incorporated throughout the year. In total, the course included fourteen in-class experiences with simulation-based inference, including multiple opportunities for groups of students to design their own simulations. More details about these activities are provided in Appendix E.

Data Collection

The data for this study were collected from AP Statistics students taught by the author in two different years. In the pilot study, seven students – selected to represent a range of statistical understanding – were interviewed individually in the weeks following the AP Statistics exam, and these interviews were audio-recorded and transcribed. Two years later, a more comprehensive dissertation study was conducted with a second class of AP Statistics students. In addition to individual interviews, data collection in this phase included student responses to targeted formative assessments, exam items, and survey items; daily field notes and journal entries written by the teacher-researcher; and transcripts and written work from group interviews.

Individual interviews

The individual interview tasks prompted students to conduct hypothesis tests to draw conclusions about the results of research studies. These tasks are reproduced in Figure 4-1 (task 1 and task 2). All students were asked to apply two different methods – a traditional test and a simulation-based test – to a single given context. All tools necessary to carry out the two approaches were provided to students; these included chance devices (e.g., coins, dice, and cards), computer applets, and graphing calculators. As students worked, they were encouraged to think aloud and provide any relevant visual representations. After carrying out both approaches, students were asked to compare and contrast the two approaches and describe any connections they saw between them. In addition to the transcripts of the task-based interviews, students'

Interview Tasks

Task 1: Helper vs. Hinderer

In a study reported in *Nature*, researchers investigated whether infants take into account an individual’s actions towards others in evaluating that individual as appealing or aversive, perhaps laying the foundation for social interaction. In one component of the study, 10-month-old infants were shown a “climber” character that could not make it up a hill in two tries. Then they were shown two scenarios for the climber’s next try, one where the climber was pushed to the top of the hill by another character (“helper”) and one where the climber was pushed back down the hill by another character (“hinderer”). The infant was alternatively shown these two scenarios several times. Then the child was presented with the two characters from the video (the helper and the hinderer) and asked to pick one to play with. The researchers found that 14 of the 16 infants chose the helper over the hinderer.

Task 2: Oil and Blood Pressure

In a study reported in the *New England Journal of Medicine*, researchers investigated whether fish oil can help reduce blood pressure. 14 males with high blood pressure were recruited and randomly assigned to one of two treatments. The first treatment was a four-week diet that included fish oil, and the second was a four-week diet that included regular oil. At the end of the four weeks, each volunteer’s blood pressure was measured again and the reduction in diastolic blood pressure was recorded. The results of this study are shown below. Note that a negative value means that the subject blood pressure increased.

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

Task 3: Response Bias

In Chapter 4 we learned how characteristics of an interviewer can lead to response bias. Two AP Statistics students decided to investigate this issue. They speculated that students would be more likely to identify as feminists if asked by a female interviewer. A sample of 60 male high school students were asked, “Are you a feminist?” Half were randomly assigned to a male interviewer and half were randomly assigned to a female interviewer. Of the 30 asked by a male interviewer, 11 responded, “Yes.” Of the 30 asked by a female interviewer, 15 responded, “Yes.

Figure 4-1. Inference tasks for individual and group interviews.

written work was collected. In total, the fourteen individual interviews were conducted – seven from each class.

Group interviews

Additionally, all eleven students enrolled in the second cohort were invited to participate in group interviews. Ten of these students were interviewed in pairs. (One was unable to participate because of absence.) Similar to the structure of the individual interviews, students were given the results of a study and were asked to work together to decide if the study provided convincing evidence. The task is reproduced in Figure 4-1 (task 3). Schoenfeld (1985) suggests that interviews with multiple students produce rich data for investigating students' problem-solving processes. Multi-person protocols ease the pressure to "produce something mathematical for the researcher," thus eliciting more natural responses (Schoenfeld, 1985, p. 178). Further, discussions among students makes the reasoning behind their decisions more visible (Schoenfeld, 1985).

Student work

Formative assessments and exam items were intended to assess students' developing understanding of inference and associated concepts, such as sampling distributions and p-values. Additionally, some items prompted students to reflect on their use of models and representations or draw connections between inferential concepts. These assessment items are included in Appendix D.

Teacher reflections

Immediately after each lesson, the teacher-researcher wrote a journal entry to record her observations of student thinking. These journal entries were based on brief, informal field notes taken during the lesson. In addition to providing context, these journal entries were intended to capture observations of classroom activity that may inform the research questions.

Data Analysis

Data analysis consisted of a process of systematic coding in multiple phases, according to the guidelines for grounded theory presented by Charmaz (2014). These guidelines provided a systematic yet flexible way to study the emerging data through constant comparisons among data, codes, and categories rather than *a priori* theory. In the initial coding phase, each segment of data was assigned a concrete and descriptive code intended to reflect students' actions. After comparing these initial codes to the data and looking for patterns in the codes across interviews, the researcher inductively constructed a set of focused codes to identify common errors. Data coded for common errors was then subjected to systematic comparisons. First, data assigned the same code were compared across participants. Second, data coded for common errors were compared to other work produced by the same participant. These comparisons aimed to contextualize the errors to better understand the conceptions they represent. Ultimately, these comparisons revealed error patterns, which provided the basis for a new conceptualization of the logic of inference. This framework is presented in the next section.

Conceptualizing the Logic of Inference

This study was informed by a by a *models and modeling* theoretical perspective, which begins with two assumptions:

> (a) People interpret their experiences using models.
> (b) These models consist of conceptual systems that are expressed using a variety of interacting media (concrete materials, written symbols, spoken language) for constructing, describing, explaining, manipulating, or controlling systems… (Lesh & Doerr, 2003a, p. 536)

Researchers working in this perspective are often interested in models that correspond to the “real world”, but the models and modeling perspective offers ontological flexibility; this is, the perspective does not make any claims about the nature of reality: “Following the pragmatists … models and modeling does not concern itself with truth. Models are adopted or rejected because they are useful...” (Lesh & Doerr, 2003a, p. 538). This pragmatic ontology is well-suited to the modeling process relevant to statistical inference; as will be discussed later in this section, the models used for inference represent the null hypothesis, not a best-guess representation of reality.

Statistical inference involves questioning whether an observed result is surprising given a particular expectation or claim (Zieffler, Garfield, delMas, & Reading, 2008). An observed result that was unlikely to occur by chance under the given claim provides evidence against that claim. Thus, statistical inference employs a type of reasoning known as *modus tollens*, which tends to be difficult for students (delMas, 2004): Suppose statement *p* implies statement *q*. If *q* is not true, then it follows that *p* is not true.

Formal inferential reasoning requires “an understanding of the interconnections between an underlying theory or hypothesis that is to be tested; a sample of data that can be examined; and a distribution of a statistic for all possible samples under the assumption that the theory or hypothesis is true” (Zieffler et al., 2008, p. 45). Note that coordination of the components mentioned by Zieffler et al. (2008) requires shifting between two different perspectives. The sample data were produced by a randomized process in the real world. On the other hand, the distribution of the statistic, or the sampling distribution, is constructed based on the assumption of a hypothesis that may or may not be true. The hypothesized model was constructed, not as a best-guess representation of the real world, but as a model whose rejection might have explanatory power in the real world. The two perspectives – the real world and the hypothetical – are linked by the hypothesis that is being tested – the null hypothesis.

The goal of statistical inference is to use a statistic calculated from a particular sample or experiment in the real world to draw inferences, often about a larger population or an underlying causal relationship. Because statistics vary, the observed data is not expected to reflect the true parameter or true relationship perfectly; consequently statistical inference is more complex than *modus tollens* logic in a deterministic scenario. In order to account for imperfect correspondence between the sample statistic and the population parameter, students must consider the distribution of statistics that could be generated by the hypothesized model. Thus, statistical inference methods entail consideration of three levels: the true relationship or population distribution, the distribution of a single sample, and the distribution of statistics calculated from multiple samples.

As illustrated in Figure 4-2, the logic of inference requires coordination of two perspectives at three distinct levels. Consider the experiment described in task two, which tested the effect of fish oil on blood pressure. On average, men in the fish oil group saw larger reductions in blood pressure than men in the regular oil group. However these data may not perfectly reflect the unknown, real-world relationship between fish oil and blood pressure. It’s possible that fish oil has no effect on blood pressure and the difference in sample means for the two groups was due to random assignment alone; this possibility is called the null hypothesis.

In order to use real-world empirical results to evaluate the null hypothesis, students must consider possible empirical results that could be generated by the hypothesized model, shifting to a hypothetical perspective. A model is specified to approximate the variability in outcomes that would occur due to randomization alone if the null hypothesis were true. For example, in the fish oil experiment, random assignment to groups can be simulated using cards. The improvement scores are written on cards, then cards from both groups are shuffled together. The hypothetical dotplot[1] in Figure 4-2, shows the empirical results of one such simulated trial. Although this model uses real-world data, it requires a hypothetical perspective, since it assumes that the fish oil and regular oil had no effect and reductions in blood pressure were determined by factors unrelated to treatment.

The process is then repeated many times to create distribution of summary statistics that shows which outcomes are typical under the specified model. The hypothetical histogram in Figure 4-2 shows a distribution of differences of sample means, calculated from 1000 simulated samples. The distribution is centered at zero, reflecting the assumption of the null hypothesis.

Note that the term *simulate* is used differently here than it is used in other disciplines like science. In probability and statistics instruction, simulations tend to require a hypothetical perspective, since they begin with the assumption that a population parameter is known, as it would rarely be in the real-world. However, in other disciplines, simulation models may be constructed as a best-guess representation of reality and used to predict what would happen in the real-world. Thus, novices may not

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

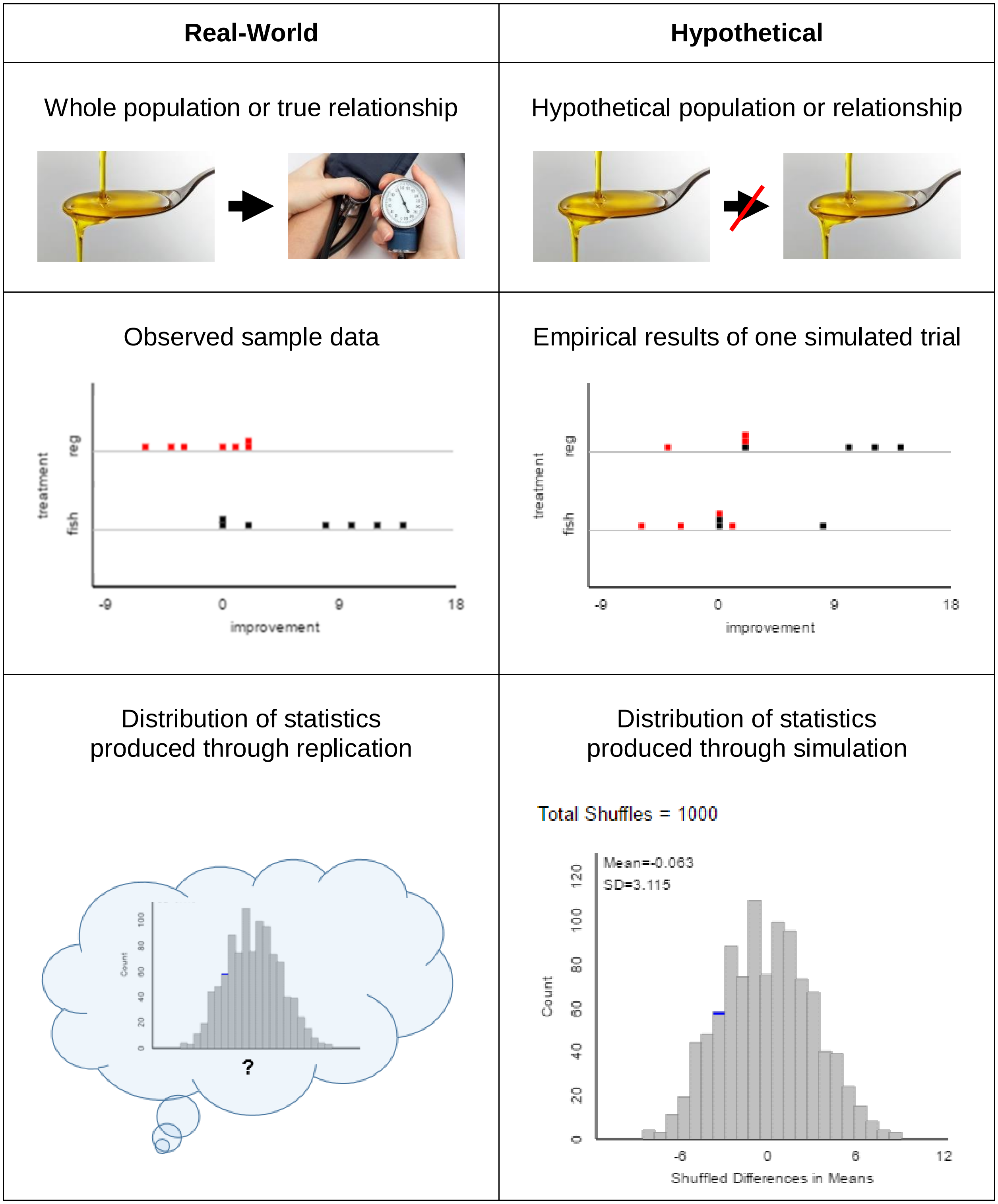

Real-World
Hypothetical
Whole population or true relationship
Hypothetical population or relationship
Observed sample data
treatment
reg
fish
-9
0
9
18
improvement
Empirical results of one simulated trial
treatment
reg
fish
-9
0
9
18
improvement
Distribution of statistics produced through replication
Count
0
20
40
60
80
100
?
Distribution of statistics produced through simulation
Total Shuffles = 1000
Mean=-0.063
SD=3.115
Count
0
20
40
60
80
100
120
-6
0
6
12
Shuffled Differences in Means

Figure 4-2. Levels and dimensions of inferential reasoning.

appreciate the subtle shift to the hypothetical perspective that occurs during statistical inference.

In the final step, students evaluate the strength of evidence by comparing the observed statistic in the real world to the distribution of statistics produced under the assumption of the null hypothesis. In the fish oil example, the observed difference of means of 7.714 falls in the tail of the distribution. Because a difference of 7.714 would be very unlikely to occur by random assignment if the treatment had no effect, we reject the hypothesized model for the relationship between fish oil and regular oil.

Of the six elements shown in the Figure 4-2, only one – the distribution of sample data – is observable in the real world. The true, real-world relationship between fish oil and blood pressure is never known. (Rejecting a "by chance alone" explanation does not deny that chance had some effect.) Further, we never see a distribution of statistics produced through real-world replication of the study, but as described in the next section, the concept of such a distribution can be a source of confusion for students.

Common Errors and Underlying Conceptions

Many of the common errors associated with simulation-based inference are associated with the challenge of coordinating multiple perspectives and levels simultaneously. This section describes errors that arise when student conflate or fail to make appropriate connections among the components shown in Figure 4-2.

Distinguishing Samples and Sampling Distributions

Some statistics educators have found that students struggle to distinguish the three levels: the parent population, the distribution of a single sample, and the sampling distribution. Saldanha and Thompson[2] (2002) investigated students' reasoning about samples, sampling distributions, and margins of error in a high school statistics class. As part of the teaching experiment, students viewed computer simulations of repeated random sampling from a population. The study found that most students were not able to relate individual sample outcomes to a distribution of outcomes in ways that supported inferential reasoning; for example, some students interpreted probabilities calculated from a simulated sampling distribution in terms of the original experimental units (people) rather than simulated statistics (sample proportions) (Saldanha & Thompson, 2002).

This challenge is not unique to high school students; Rossman and Chance (2014) note that some college students also struggle to identify the observational units in a randomization distribution. Ideally, students should recognize the units in a simulated sampling distribution as both a repetition of the random process and a simulated value of the statistic (Rossman & Chance, 2014).

In the present study, no students explicitly referred to simulated statistics at the sampling distribution level as if they were sample data while using the applet to carry out simulation-based inference. However, confusion of samples with sampling distribution was made visible through confusion of the sample size with the number of trials. For example, when asked why we should repeat the simulate many times, students' answers were often ambiguous; advantages like less variability or more

[2] Saldanha and Thompson (2002) describe the same challenges related to the "multi-tiered sampling process", but they define their levels differently than levels were defined in the previous section. Namely, they define three levels in terms of *processes*: randomly selecting individuals and recording a statistic, repeating the process to accumulate statistics, and finding the proportion of statistics beyond a threshold value (p. 261).

accuracy could be interpreted as advantages of large sample sizes rather than a large number of trials. In other cases, the confusion was more clear-cut. In the incident below, Isabella and Devon compare the samples size to the number of trials as they reason about the slightly different p-value produced by traditional and simulation-based methods.

Isabella: [In the traditional method], we also had a smaller sample than this.

Devon: Oh, exactly.

Catherine: Wait, what?

Isabella: We had – like our [original] sample was smaller than that sample [of simulated statistics shown on the applet].

Devon: Yea, so [the p-value calculated using the applet] could be the true p-value.

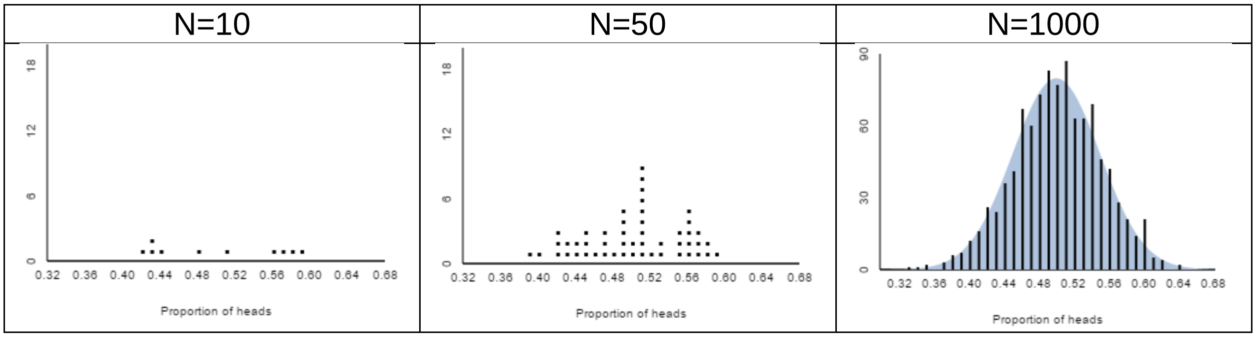


Figure 4-3. Simulated sampling distribution of a sample proportion.

Students in this study – who had encountered both traditional and simulation-based inference in instruction – also experienced confusion about whether the Normality condition was related to sample size or the number of shuffles. This conception was first observed in class, with students using Normality as a kind of stopping condition as they added to the number of simulated trials. Consider the simulated sampling distribution of a sample proportion, where each trial includes 100 flips of a fair coin. Because the sample size is large, the Normal distribution provides an appropriate model for the sampling distribution; however, as shown in Figure 4-3, that only becomes evident when the number of simulated trials is also large. In the individual interviews, a few students expressed the belief that a large number of simulated trials would satisfy the Normal condition, somehow compensating for a small sample size.

Transitioning from the Sample Level to the Sampling Distribution Level

Although no students explicitly confused the units in the sample and the units in the sampling distribution while using an applet, some students struggled to envision the multi-level nature of inference when the applet was not open. Specifically, they struggled to transition from the sample level to the sampling distribution level. This issue was visible through student work in class and on exams as well as through task-based interviews.

For example, in her individual interview, Eva designed a simulation using cards to re-randomize improvement scores, and the interviewer asked her what she would do after dealing the cards into two groups. Eva pointed to the graph of the sample data while describing her plan to repeat the process many times and graph the results.

Eva: I mean I would probably – you would put it on the graph wouldn't you? You put like each – like this (referring to graph of sample data on sheet) … you would put that on the graph. You would just graph what you got a bunch of times.

Catherine: So you would have a bunch of graphs that look like this?

Eva: No, it would be the same graph.

Catherine: The same graph. So like what would – each dot on the graph, what would each dot be?

Eva: Each dot would be the improvement score.

Catherine: Improvement score for a single person?

Eva: Yeah.

At this point, Eva recognized the need to repeat the randomization process many times, but she did not envision the next level where the units on the graph are simulated statistics. However, later in the interview when the applet was introduced, Eva was able to reason with the graphical representation of simulated statistics, describing the units as a difference of means in context.

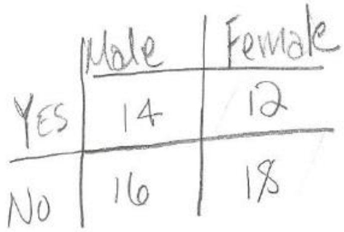
Zieffler, delMas, Garfield, and Brown (2014) reported a similar phenomenon among students enrolled in the CATALST course – a college level course that uses simulation in TinkerPlots to carry out statistical inference. One student designed a model and described simulation of a single trial but was initially unable to explain how this information would be used. However, when using TinkerPlots software, the same student was able to complete the process and draw a conclusion about whether the observed result was surprising (Zieffler, et al., 2014).

In the individual and group interviews collected in this study, difficulty transitioning to the sampling distribution level sometimes led to inappropriate comparisons across perspectives at the sample level. For example, in her group interview, Eva again struggled to move from the level of sample to the level of sampling distribution. As Eva worked with her partner Ryan, they considered two approaches to compare the observed sample data with the results of a single simulated trial. The data from the original study (task 3) are reproduced in Table 4-1.

Table 4-1. Data for group interview task.

Feminist?	Male Interviewer	Female Interviewer	Total
Yes	11	15	26
No	19	15	34
Total	30	30	60

In the incident below, Eva and Ryan had just simulated one trial using their physical model; their simulated results are shown in Figure 4-4.



	Male	Female
YES	14	12
No	16	18

Figure 4-4. Student work.

Ryan: Now we compare it to our results here. That would be our observed counts. I mean, not our observed counts – that would be our expected counts.

Eva: We don't do another chi-square?

Ryan: No… When you simulate, I'm pretty sure you just – because a chi-square test is an inference test. This is a simulation test.

Eva: I trust you.

Ryan: So does the proportion – so the proportion of people who said yes with a male interviewer was actually lower than who would have, if that makes sense.

Eva: What this?

Ryan: And then… By chance – if it was by chance then the people who said yes when they were interviewed by a female would have been 12, but the ones who did say yes was 15.

Ryan treated the simulated counts as expected counts, because their model was designed to see what happens by chance. However, at this point in the interview, he did not acknowledge the need for a distribution of simulated statistics; instead, he used a single simulated sample as an indication of what would happen just by chance. Taking a different approach, Eva suggested that the simulated data be used as the basis for a chi-square test. This approach was common among the students in this study, who were exposed to both traditional and simulation-based tests in class. Students' justifications highlight the challenge of coordinating real-world and hypothetical perspectives. Consider the following exchange from Libby's individual interview, where she explains how to use the simulated data:

Libby: Yea, so I plug this into L1 and this into L2 [on the calculator] and I would use a two-sample t-test and basically see what my t and p-value are.

Catherine: So you'd do another t-test but on your simulated data?

Libby: Yes.

Catherine: What would it tell you? Like pretend you got a p-value of 0.3?

Libby: It would tell me that this is – like this outcome with this data – if it was 0.3 – is way more likely to occur just by chance.

Catherine: Right, so this did occur just by chance, right?

Libby: Yea, right, right. But the calculator doesn't know that.

Distinguishing Simulation and Replication

Another set of common errors stem from students imagining a distribution produced through replication. In the real world, we never see a distribution of statistics estimated from a large number of studies, but it is common for students to conflate simulation with replication – shifting from a hypothetical perspective to a real-world perspective. This conception can manifest as a number of different errors.

One problem that can arise is the belief that multiple samples are always necessary (Hodgson & Burke, 2000). In an activity intended to develop understanding of the Central Limit Theorem, students in an introductory statistic course repeatedly selected samples from a given parent population and constructed histograms of the resulting sample means. In an assessment given immediately after the activity, one-third of students expressed the belief that multiple samples are necessary for valid statistical inference. Although the activity maintained a hypothetical perspective, where the parent population is somehow known, some students mistook the process for "a real-world strategy for finding a population parameter" (Hodgson & Burke, 2000, p. 94).

This issue also arises in simulation-based inference and may persist, even in courses that devote extensive time to simulation: “Some mistakenly believe that simulation aims to provide replication of the research study, in order to strengthen the findings through replication (Rossman & Chance, 2014, p. 218). This error appeared several times in the present study. Unlike the error described earlier, these students do not see simulated samples as a way to increase the sample size but as a way to replicate the entire study. Thus, students may believe that errors in the first study can be corrected in the replicated studies. For example, Laura obtained substantially different p-values from the two approaches because of a calculation error; she offered the follow explanation for the discrepancy:

Laura: In the [traditional test], this was just one sample, so maybe since it was just one sample, there might have been factors that affected the fish oil and the regular oil, but since [the simulation-based test] was over time, maybe this kind of eliminates more of those factors. Or it – or yea, so it eliminates different confounding factors.

Conflating simulation and replication can also lead to misapplications of the logic of inference, which prevent students from rejecting the null hypothesis, regardless of the data. Gould et al., (2010, p. 4) report that often “the null distribution of the test statistic is seen as the ‘real’ distribution, and students reason that because the distribution is centered at 0, the null hypothesis cannot be rejected.” A closely related error is to count how many samples are “more extreme” than the center of the sampling distribution. Maria made these errors as she considered whether 14 out of 16 babies choosing the helper toy provided convincing evidence of a genuine preference over the hinderer toy in task 1 (testing the alternative hypothesis $p_{helper} > 0.5$.)

Maria: I think you would take all of the ones – all of the numbers that are higher than 8, but since you did so many trials that seems like a lot, but it doesn't look like there is convincing evidence, because it's centered at 8.

Other students who conflate simulation and replication do find the proportion of samples more extreme than the observed data. However, if that proportion is small, students reason that the original data must have been an outlier or a fluke – a result that is unlikely be replicated. Some students who follow this line of reasoning still employ "by chance" language. Notice how Anthony employs statistical terms as he reasons about the helper vs. hinderer study in task 1.

Anthony: So again, just looking to see if there's any data points at 14 and there's not any or any past it, so it would be unlikely – it would be unlikely that this would occur – so we'd say it'd be statistically significant. … Based on this data, we'd conclude that the data would have occurred by chance. Given that there's no – there is not much evidence at all for supporting that 14 of the 16 would have chosen the helper over the hinderer, because what we're seeing is there's more of an equal chance.

Discussion

Coordination of Levels and Dimensions

Saldanha and Thompson (2002) described coordination of samples and sampling distributions among students in their teaching experiment as "unstable":

> Most students experienced great difficulty conceiving the resampling process in terms of distinct levels … Their control of the coordination between the various levels of imagery was unstable; from one moment to the next their image of a number of samples (of people) seemed to easily dissolve into an image of a total number of people. (Saldanha & Thompson, 2002, p. 264)

Disappointed with the results of the teaching experiment, they speculated that the simulation activities "were of such a complexity so as to essentially overshadow ideas of sampling variability" (Saldanha & Thompson, 2002, p. 268) While the present study also identified difficulty with coordination between levels, students often resolved these issues and went on to reason about the statistical concepts under study.

The present study also found that students' coordination of real-world and hypothetical perspectives was "unstable" or transitory. For example, few students persistently reasoned that simulation was equivalent to replication. More often, the transcripts show students struggling to reconcile the hypothesized model with a conception of real-world replication. These findings are consistent with Lesh and Doerr's description of students' developing conceptual systems as "less like rigid and stable worlds than they are like … shifting collections of tectonic plates" (Lesh & Doerr, 2003b, p. 18).

In some cases, students were able to design a simulation and justify their choice of physical model based on the null assumption, but later interpreted the empirical results from a real-world perspective. As illustrated above, Anthony treated the simulated distribution as a representation of real-world replication. However, earlier in the interview, he chose a coin as a physical model without prompting from the researcher, justifying his choice by reasoning that a coin would represent the "same chance of being picked by each infant." These inconsistent interpretations of a system may be associated with the use of representational media that emphasize different aspects of the underlying systems (Johnson & Lesh, 2003; Lesh & Doerr, 2003b).

At one point in the individual interview, Eva's reasoning was similar to Anthony's. Because the observed statistic rarely occurred in the simulated distribution, she reasoned it must be a "just by chance guy" – an outlier or a fluke. However, in the same breath, she acknowledged her expectation that the distribution would be centered at 8, because it was based on a fair coin; that is, she acknowledged the null assumption.

Eva: Ok, so it's looking like 14, which is what we had, doesn't have that many – it's not like it – it's not skewed over there. This is a just by chance guy. And it happened – it's centered at 8, so you know, it's half. It's what you would think with a fair coin.

Catherine: So can you explain that other part? You said something about skew and a just by chance guy?

Eva: It's Normal, so it's centered at what I drew before with the… It's centered at 8, so that's like the half mark of how many babies in the thing that would pick the helper. And because it's centered at 8, that means that there is no difference between picking the helper or the hinderer.

For Eva, this conflict of conceptions was temporary. First, she noticed that she had come to a different conclusion than the one she drew from the traditional test. When the interviewer pointed out that that she hadn't used the original data, she remembered that the applet could be used to count samples as extreme as the sample data.

Eva: (using the applet to count samples more extreme than 14 out of 16) That is so small.

Catherine: 3 out of 1000 – what does that tell you?

Eva: That getting 14 out of the 16 to choose the helper is very unlikely.

Catherine: So if the question is, "Is this convincing evidence?"…

Eva: Then yes, because you would – based off of this you would suspect that – if it was just by chance you expect that only 8 out of the 16 babies that were in the study would choose the helper, but because this is so unlikely, then you know that – because it's unlikely, it's not likely to happen just by chance. So if there's that small of a percentage chance that you're gonna get 14 babies to pick the helper then you know that your one little trial study thing is significant.

Not only did Eva transition to a more productive conception of the simulated sampling distribution, but she was able contrast her new and previously held conceptions. Her insights allowed the researcher to better understand the work of other students who conflated simulation and replication, making it possible to link common errors not previously seen as similar.

The results of this study provide context for an inconsistency reported by Chance and McGaughey (2014): students seem to understand significance tests as a way to decide whether observed results could have happened by chance alone, yet they do not appreciate the role of the null hypothesis for estimation and interpretation of the p-value. First, this study substantiates a warning from Pfannkuch (2005) that the term “by chance alone” is not universally understood; in fact, students may incorporate this language into an alternative logic of inference. Second, students struggle to simultaneously coordinate the multiple perspectives and levels that compose the logic of inference. That is, they may recognize the foundational role of the null assumption at some points in the process but not others.

Omnipresence of Uncertainty

Statistics is often described as a set of tools for dealing with the “omnipresence of variability” (Cobb & Moore, 1997). Statisticians must account for variability from innumerable sources, including variability among individuals in a population, purposeful variation of conditions in an experiment, and particularly relevant for this study, variability in statistics due to random sampling and random assignment. The omnipresence of sampling variability results in omnipresence of uncertainty: although we make decisions about whether to reject the null hypothesis, the veracity of this link between the real-world and hypothetical perspectives is never known.

The distinction between making a decision and proving a hypothesis is subtle and often difficult for students. In particular, it may be difficult for students to accept uncertainty when inference requires an assumption that the parameter is known (from a hypothetical perspective). Further, some statistical definitions seem to presume that the true parameter is knowable. For example, type I error is defined as the probability of rejecting the null hypothesis when the null is really true. One student objected to this definition, because we can never prove the truth of the null hypothesis. The challenge of accepting uncertainty from a real-world perspective while assuming truth is knowable from a hypothetical perspective cannot be ignored.

Two Approaches to Inference

Some of the conceptions associated with errors in simulation-based inference also arise in courses that employ traditional inference alone. Whether constructed empirically or theoretically, sampling distributions always require a multi-level scheme that distinguishes between the population distribution, the distribution of a single sample, and the distribution of statistics calculated from multiple samples. Further, the logic of inference – which requires coordination of the real-world and hypothetical perspectives – remains unchanged across inferential approaches. As discussed in Chapter 3, simulation-based inference has the potential to make student thinking visible, so conceptions that lead to errors in simulation-based inference may go unnoticed in traditional inference. In short, the results presented do not provide a basis for rejecting simulation-based inference in favor of traditional inference, but they do serve as reminder that simulation-based methods are not a panacea.

Some of the errors described in this article are specific to courses that employ *both* traditional and simulation-based methods to introduce the logic of inference. It is worth noting that some of the proposed advantages of simulation-based inference – e.g., avoiding mathematical formulas and theoretical sampling distributions – do not apply to courses that use simulations in addition to traditional tests. In these courses, empirically-derived sampling distributions are yet another set of models to represent outcomes under the null hypothesis. As described above, additional models can lead to confusion as students make inappropriate connections between traditional and simulation-based approaches. However, these inappropriate connections must be weighed against the productive connections students make between approaches; this is discussed in more detail in Chapter 5.

Summary

Although simulation-based inference offers a number of proposed advantages over traditional inference alone, simulation methods give rise to a number of common errors. These errors can be described largely in terms of two challenges. First, students struggle to coordinate the multi-level scheme, which includes the population or true underlying relationship, the distribution of single sample, and the distribution of statistics collected from multiple samples. Second, students struggle to coordinate two perspectives: the real-world where the sample data was collected and the hypothetical perspective where the null hypothesis is assumed to be true.

The logic of inference always requires coordination across multiple levels and perspectives, so these challenges are not unique to simulation-based inference. Further, the students in this study were exposed to both traditional and simulation-based methods in instructions, so no conclusions can be drawn about the causes of their conceptions. However, the results of this study suggest that students demonstrate their conceptions in distinctive ways as they reason about inference tasks using simulation-based methods. Awareness of common errors provides an opportunity for teachers to recognize and challenge students' emerging conceptions of inference.

CHAPTER 5
TEACHING FOR CONNECTED UNDERSTANDING OF INFERENCE

Many introductory statistics courses emphasize statistical inference as an important objective of the course. Significance testing is a widely used data analysis tool (Nickerson, 2000), and although no introductory course can include all hypothesis tests, “a conceptual understanding of the p-value and statistical significance opens the door to a wide array of statistical procedures that utilize this inferential logic” (Lane-Getaz, 2007, p. 10). However, statistical significance and p-values are commonly misunderstood; specifically, many students in introductory statistics courses understand p-values as a tool for making decisions about the null hypothesis or a way to quantify the strength of evidence but lack an integrated conceptual understanding of what the p-value represents (Aquilonious & Brenner, 2015; Holcomb, Chance, Rossman, & Cobb, 2010; Taylor & Doehler, 2015).

Some statistics educators (e.g., Chance & Rossman, 2006; Cobb, 2007; delMas et al., 1999; Pfannkuch, 2005) believe that simulations have the potential to develop a deeper conceptual understanding of statistical significance and p-values. Proponents have argued that these methods require less prerequisite knowledge, generalize easily to a large number of settings, incorporate modern computing power in a meaningful way, and support conceptual understanding of inference (Chance & Rossman, 2006; Cobb, 2007; Holcomb, Chance, Rossman, Tietjen, et al., 2010). Today, simulation-based inference methods are increasingly common in introductory statistics courses (ASA, 2016; Rossman & Chance, 2014): some instructors have incorporated a few activities or modules, while others have completely reconceptualized their courses with simulation-based inference as the cornerstone.

Developers of curricula that employ simulation-based inference as the primary means of teaching inference have published rich descriptions of their inferential instruction (Garfield et al., 2012; Lock et al., 2014; Tintle et al., 2011). Further, many individual simulation-based inference activities have been shared through conference presentations and the practitioner literature. However, relatively little has been written about traditional courses that regularly incorporate simulation-based inference as a complement to theory-based methods. This article provides a detailed description of an AP Statistics course that used traditional inference as the primary means of data analysis but also included numerous experiences with simulation-based inference to develop concepts.

Teachers who complement traditional inference with simulation-based methods hope to expand students' capacity to reason about inference. The potential of this pedagogical approach seems to rest on the assumption that students make productive connections between the two models and representational systems. This chapter presents an investigation of that assumption, discussing which connections students tend to make readily and which require focused instruction.

Complementing Traditional Instruction with Simulation-Based Inference

This study is situated in the context of an AP Statistics class at P.K. Yonge (PKY) Developmental Research School. In addition to the prescribed AP Statistics curriculum, which relies on traditional inference methods (College Board, 2010), the course taught at PKY regularly incorporated simulation-based inference methods. However, students in this course had considerably more experience with traditional, theory-based methods by the end of the school year. This section describes how the course used simulation-based inference activities as a complement to traditional inference in instruction.

Consider the following inference task, which necessitates an inference method to determine the statistical significance of experimental results. This task is based on an activity in the teacher's edition of *The Practice of Statistics* (Starnes et al., 2013).

> In a study reported in the *New England Journal of Medicine*, researchers investigated whether fish oil can help reduce blood pressure. 14 males with high blood pressure were recruited and randomly assigned to one of two treatments. The first treatment was a four-week diet that included fish oil, and the second was a four-week diet that included regular oil. At the end of the four weeks, each volunteer's blood pressure was measured again and the reduction in diastolic blood pressure was recorded. The results of this study are shown below. Note that a negative value means that the subject blood pressure increased.

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

This example will be used to illustrate the pedagogy specific to this course and the approach to inference embodied in the AP Statistics curriculum more generally.

Traditional Inference Instruction

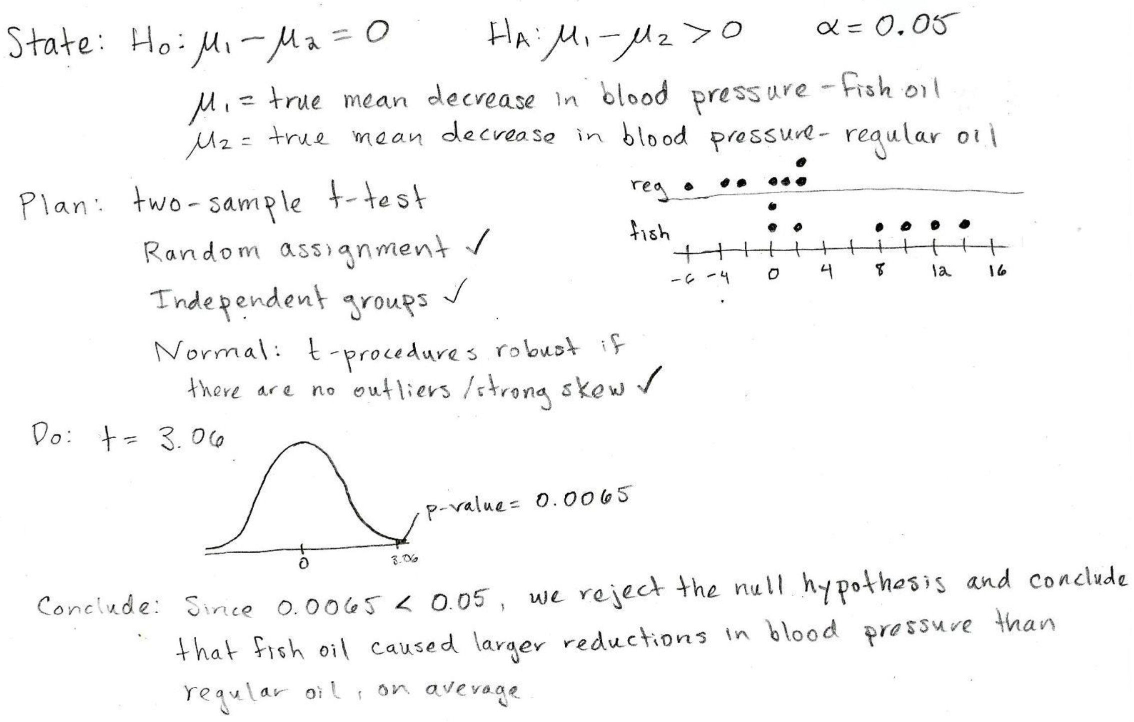
The AP Statistics course description includes four major topics (College Board, 2010): data analysis and exploration (20-30%), study design (10-15%), probability and simulation (20-30%), and statistical inference (30-40%). Because traditional inference requires substantial prerequisite knowledge, including knowledge of probability and theoretical sampling distributions, traditional inference is typically taught in the final third of the course (Malone et al., 2010). The course textbook, *The Practice of Statistics* (Starnes et al., 2013), adheres to this pattern, covering traditional inference at the end of the year. The AP Statistics curriculum includes nine tests of significance (College Board, 2010): large-sample test for a proportion; large-sample test for a difference between two proportions; test for a mean; test for a difference between two means (unpaired and paired); chi-square test for goodness of fit, homogeneity of proportions, and independence; and test for the slope of a least-squares regression line.

To help students organize statistical problems, *The Practice of Statistics* (Starnes et al., 2013) uses the same four-step process throughout the text. The four-step process, as applied to all significance tests, is as follows:

> - State: What *hypotheses* do you want to test, and at what significance level? Define any *parameters* you use.
> - Plan: Choose the appropriate inference *method*. Check *conditions*.
> - Do: If the conditions are met, perform calculations.
>   - Compute the test statistic.
>   - Find the ***P*-value**.
> - Conclude: *Interpret* the result of your test in the context of the problem. (Starnes et al., 2013, p. 552, emphasis in original)

This four-step process is applied to all traditional tests of significance in the course, and over a period of several months, students apply the process times in class, on assignments, and on tests.

Following this framework, hypotheses are stated using formal inscriptions. The appropriateness of a particular inference method is determined by checking a list of conditions. Calculations are introduced through formulas, but later carried out largely by dedicated inference functions in the TI-84 Plus calculator. Conclusions based on the p-value are always interpreted in context. For example, the written work in figure illustrates how a high-achieving student in the class might approach the fish oil experiment described above.



State: $H_0: \mu_1 - \mu_2 = 0$ $H_A: \mu_1 - \mu_2 > 0$ $\alpha = 0.05$

μ_1 = true mean decrease in blood pressure – fish oil

μ_2 = true mean decrease in blood pressure – regular oil

Plan: two-sample t-test

Random assignment ✓

Independent groups ✓

Normal: t-procedures robust if there are no outliers / strong skew ✓

Do: $t = 3.06$

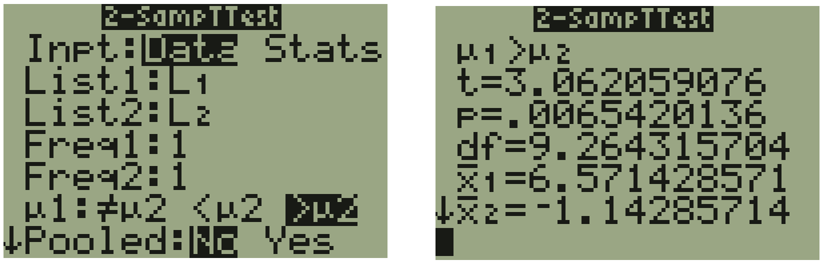
p-value = 0.0065

Conclude: Since $0.0065 < 0.05$, we reject the null hypothesis and conclude that fish oil caused larger reductions in blood pressure than regular oil, on average.

Figure 5-1. Traditional inference using State-Plan-Do-Conclude framework.

Cobb (2007, p. 2) points out, evaluating “the fit between model and reality” can be technically complicated, even in this seemingly simple case of a difference between two means. Cobb (2007, p. 8) also disputes the use of a sampling modeling to represent the outcomes of randomized experiments: “Do we want students to leave their brains behind and pretend, as we ourselves apparently pretend, that choosing at random from a large normal population is a good model for randomly assigning treatments?” However, this course does not expose students to all the details of theoretical models; the textbook (Starnes et al., 2013) streamlines the process of checking model fit by including lists of conditions for each test.

In this course, the calculations for a given hypothesis test are introduced using formulas for the test statistic and cumulative density functions in the TI-84 Plus calculator. Probability tables are never used in this course. After using these tools a few times, students are introduced to functions on the TI-84 Plus that calculate test statistics and p-values using summary statistics or raw data as inputs. Note that the calculator's inference functions allow students to calculate the test statistic and p-value, as required for the AP Statistics exam, without using formulas or creating visual representations of the sampling distribution.



```
2-SampTTest
Inpt:Data Stats
List1:L1
List2:L2
Freq1:1
Freq2:1
μ1:≠μ2 <μ2 >μ2
↓Pooled:No Yes
```

```
2-SampTTest
μ1>μ2
t=3.062059076
p=.0065420136
df=9.264315704
x̄1=6.571428571
↓x̄2=-1.14285714
```

Figure 5-2. Calculating a p-value using a TI-84 Plus inference function.

Simulation-Based Inference Instruction

In addition to traditional inference methods, the course taught at PKY regularly incorporated simulation-based inference methods. The use of simulation-based inference to complement traditional instruction was supported by the course textbook, *The Practice of Statistics* (Starnes et al., 2013); however, the course also used activities drawn from other sources. In 2015-2016, the class included fourteen in-class simulation-based inference activities; detailed information about these activities is provided in Appendix E.

To help students recognize the unified modeling process of simulation-based inference, the teacher adopted the 3S Strategy used in the *Introduction to Statistical*

Investigations curriculum (Tintle, et al., 2013), applying the same approach to each inference task:

1. Statistic: Compute the statistic from the observed sample data.
2. Simulate: Identify a “by chance alone” explanation for the data. Repeatedly simulate values of the statistic that could have happened when the chance model is true.
3. Strength of Evidence: Consider whether the value of the observed statistic from the research study is unlikely to occur if the chance model is true. If we decide the observed statistic is unlikely to occur by chance alone, then we can conclude that the observed data provide strong evidence against the plausibility of the chance model… (Tintle, et al., 2013)

This three-step process was applied to all simulation-based tests of significance. In the following sections, the process is used to outline how simulation-based inference was taught in this course.

Statistic

The data from the original experiment can be summarized by a difference of sample means: $\bar{x}_1 - \bar{x}_2 = 6.57 - (-1.14) = 7.14$. Unlike traditional methods, this approach does not require calculation of a standardized test statistic. There are two possible explanations for the observed difference of sample means: The first is a “by chance alone” explanation; it is possible that the treatments had no effect, and the difference between groups is due to random assignment. The second is that fish oil caused larger reductions in blood pressure than regular oil, on average.

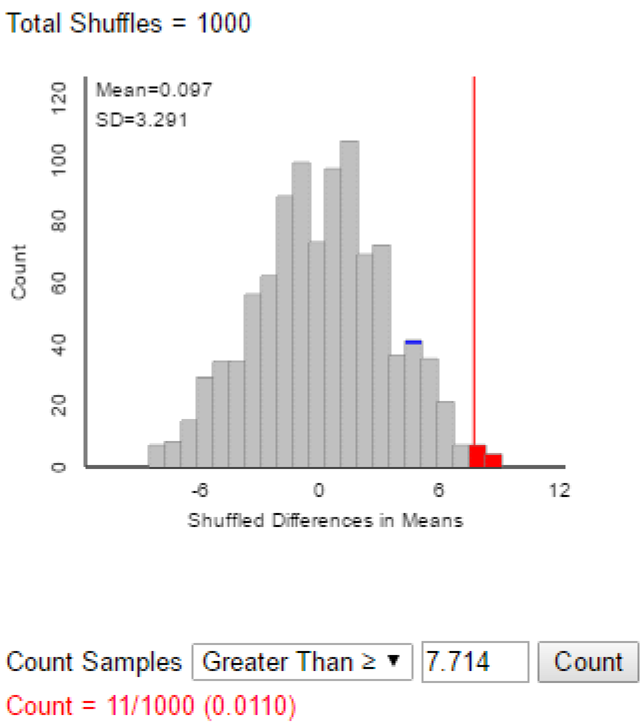
Simulate

Random assignment to groups can be modeled using cards. First we write the improvement scores on blank cards. Then cards from both groups are shuffled together, assuming that the subjects’ improvement scores were determined in advance by factors unrelated to treatment. Then the cards are dealt into two groups to mimic random assignment, and the difference in means for the two groups is recorded. We can use an applet to simulate many trials. The results of this simulation are shown in the Figure 2-3.

Students may engage with the simulation in various ways:

- Students physically carry out the simulation using cards; students may record their own simulated statistics on a class dotplot that can be used to determine the strength of evidence.
- Students use technology to carry out the simulation; for example, students may use an applet on their tablets, phones, or computers.
- Students design an appropriate simulation; the teacher may suggest a chance device (cards) or allow students to choose their own physical model.

These modes of engagement are not mutually exclusive, and this course used them in varying combinations over the course of the year. The type of student participation in the simulation step largely determines how much class time is spent on each simulation-based activity.


Total Shuffles = 1000

Mean=0.097
SD=3.291

Count Samples Greater Than ≥ 7.714

Count = 11/1000 (0.0110)

Figure 5-3. Applet to simulate random assignment.

Strength of evidence

We evaluate the strength of evidence by comparing the outcome of the original study to the distribution of outcomes produced by the model. In the simulation shown in Figure 5-3, a difference of means of 7.714 or larger occurred in 11 out of 1000 trials – an estimated p-value of 0.0011. Because a difference of 7.714 would be very unlikely to occur by random assignment if the treatment had no effect, we reject the “by chance alone” explanation, and conclude that fish oil caused larger reductions in blood pressure than regular oil, on average.

Sequencing of Topics

Notice that the example above did not require knowledge of theoretical probability distributions. Because simulation-based inference requires less prerequisite knowledge, it can be introduced early in the year. In this class, simulation-based inference was introduced on the first day, and students had many experiences with these methods before traditional inference was introduced. When traditional inference was covered during the final third of the class, simulation-based inference activities were included as part of every chapter, so students used simulations to model a variety of study designs. Aside from a brief introduction to study design during the first few weeks, this class did not substantially alter the traditional introductory statistics sequence: it began with descriptive statistics, followed by probability, and concluded with statistical inference (Malone et al., 2010). Simulation-based inference activities were inserted to coincide with related topics in the AP Statistics curriculum, namely experimental design, probability, and traditional inference.

Alternatively, a teacher might consider a spiral approach like the one employed in the *Introduction to Statistical Investigations (ISI)* (Tintle et al., 2013) curriculum. After an introduction to inferential concepts, each chapter in the ISI curriculum is devoted to a particular data scenario, e.g., comparing two proportions. Each chapter begins with descriptive statistics, followed by a simulation approach, and concludes with a theory-based approach; as the course progresses, the data scenarios become increasingly complex. This curriculum devotes less time to descriptive statistics than a typical course and does not cover formal probability rules (Tintle & Chance, 2014). Depending on the goals of the course, some teachers – in particular, AP Statistics teachers – may find it necessary to supplement the ISI instructional sequence.

Fostering Connections across Approaches

Lastly, the course taught at PKY aimsed to make the connections between simulation-based inference and traditional inference explicit. For example, traditional inference was first introduced as a modification to the 3S Strategy, where the simulation step was replaced by use of a theoretical sampling distribution. In particular, the transition began by summarizing the results of the original study with a new statistic – a standardized test statistic.

Because of the conceptual challenge of this lesson, a simple data scenario was chosen: a test of one proportion, $H_0\colon p = 0.5$. On that day, students carried out the simulation by hand using coins to represent the assumption that a p-value is calculated under the null hypothesis and based on randomness. Students were asked to calculate two different statistics – a sample proportion and a standardized z statistic – for each of their simulated samples. As a class, they created dotplots to represent the sampling distribution of each statistic using stickers. Each sticker is labeled $\hat{p}$ or z as a reminder that each dot in the distribution represents a statistic calculated from a sample.

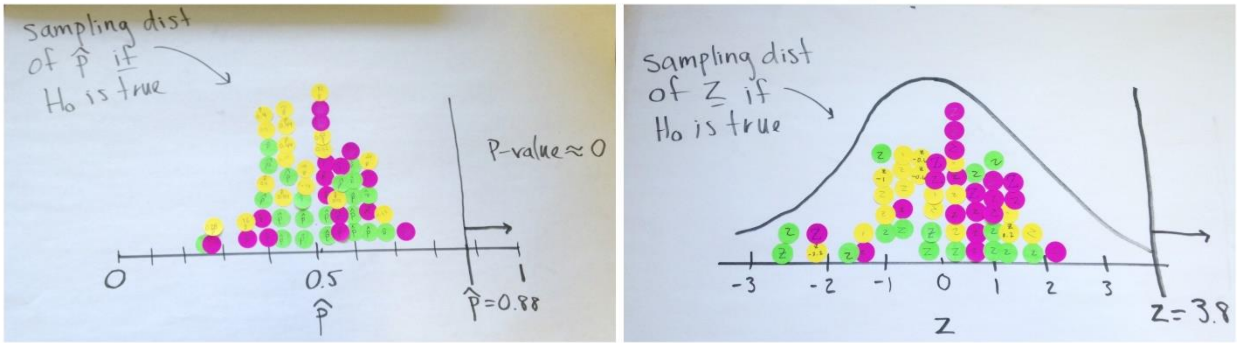


Figure 5-4. Students' simulations of sample proportions and z statistics.

The visual building up of a bell-shaped distribution centered at 0 and ranging from -3 to 3 suggests that the standard normal distribution can be used to approximate the distribution of z statistics. The teacher reminded students that they calculated the z statistics based on samples of coin flips ($p = 0.5$); thus, the z distribution represents the outcomes that would occur when the null hypothesis is true.

The p-value was estimated two different ways: first by counting the number of samples more extreme than the original data[3] and then by using a density function to estimate the area in the tail(s). These side-by-side representations emphasize that the two approaches are two ways to do the same thing – quantify the strength of evidence against the null hypothesis. This step-by-step transition can be repeated as new tests are introduced to reinforce the connections between approaches.

Lastly, this course used formative assessments that prompted students to reflect on the connections between approaches. For example, students filled out an exit ticket

[3] Figure 5-4 shows two ways a p-value was calculated from data collected in class. As it happened, the result obtained in the classroom experiment would almost never occur by chance. This can be problematic, because students can say that the outcome *never* occurred without calculating a p-value to quantify the likelihood. Teachers may want to avoid unpredictable classroom data for this important introductory lesson.

on the day z-tests were introduced (the lesson described above). Students were asked to explain how a z-test is similar to the 3S strategy and how it is different. Simple assessment like these led students to consider connections and provided the teacher with information about which connections to emphasize in instruction. A complete list of assessment items is provided in Appendix D.

Methods

Data Collection

The data for this study were collected from AP Statistics two AP Statistics classes taught by the author in two different years. In the pilot study, seven students – selected to represent a range of statistical understanding – were interviewed individually in the weeks following the AP Statistics exam, and these interviews were audio-recorded and transcribed. Two years later, a more comprehensive dissertation study was conducted with a second class of AP Statistics students. In addition to individual interviews, data collection in this phase included student responses to targeted formative assessments, exam items, and survey items; daily field notes and journal entries written by the teacher-researcher; and transcripts and written work from group interviews.

Student work. Formative assessments and exam items were intended to assess students' developing understanding of inference and associated concepts, such as sampling distributions and p-values. Additionally, some items prompt students to reflect on their use of models and representations and draw connections between inferential concepts. The assessment items used to elicit student work are included in Appendix D.

Teacher reflections. Immediately after each lesson, the teacher-researcher wrote a journal entry to record her observations of student thinking. These journal entries were based on brief, informal field notes taken during the lesson. In addition to providing context, these journal entries are intended to capture observations of classroom activity that may inform the research questions.

Individual interviews. The individual interview tasks prompted students to conduct hypothesis tests to draw conclusions about the results of research studies. These tasks are reproduced in Appendix C (task 1 and task 2). All students were asked to apply two different methods – a traditional test and a simulation-based test – to a single given context. All tools necessary to carry out the two approaches were provided to students; these included chance devices (e.g., coins, dice, and cards), computer applets, and graphing calculators. As students worked, they were encouraged to think aloud and provide any relevant visual representations. After carrying out both approaches, students were asked to compare and contrast the two approaches and describe any connections they saw between them. In addition to the transcripts of the task-based interviews, students' written work was collected. In total, the fourteen individual interviews were conducted – seven from each cohort.

Group interviews. Additionally, all eleven students enrolled in the second cohort were invited to participate in group interviews. Ten of these students were interviewed in pairs. (One was unable to participate because of absence.) Similar to the structure of the individual interviews, students were given the results of a study and were asked to work together to decide if the study provided convincing evidence. The task is reproduced in Appendix C (task 3).

Data Analysis

Data analysis consisted of a process of systematic coding in multiple phases, according to the guidelines for grounded theory presented by Charmaz (2014). These guidelines provided a systematic yet flexible way to study the emerging data through constant comparisons among data, codes, and categories rather than *a priori* theory. In the initial coding phase, each segment of data was assigned a concrete and descriptive code intended to reflect students' actions. After comparing these initial codes to the data and looking for patterns in the codes across interviews, the researcher inductively constructed a set of focused codes to identify connections across approaches.

Connections across Approaches

Testing Hypotheses

Students seem to readily accept that traditional and simulation-based approaches are ways to test the same hypotheses; of the 14 students interviewed individually, none expressed the belief that the claims being tested were different for the two approaches. Ideally, students will also make deeper connections related to the hypotheses. In particular, students should understand that both theoretical and empirical models are based on the assumption that the null hypothesis is true. Students with the highest achievement in the course often emphasized this connection. For example, when asked to describe how the two approaches were similar, Cameron mentioned the assumption of the null hypothesis first.

Cameron: I mean, in the two approaches we both ended up using the null hypothesis as our like – like our base. I don't know – that's what I refer to it as, like what we used to sample things or to model off of.

For some students, the assumption of the null hypothesis may seem abstract in traditional inference. Simulation-based inference makes the null assumption more concrete by providing a physical model for the "just by chance" explanation. Constructing a physical model requires consideration of the null hypothesis and the randomness inherent in the study design. Some students incorporated the language of modeling even when they discussed traditional tests. For example, before testing the hypothesis $H_0: p = 0.5$, Devon drew a Normal distribution, but he described it as follows:

Devon: It's like a – the null hypothesis distribution. So a coin flip distribution, I guess. Like how often it would occur by chance.

Modeling Chance Outcomes

All students interviewed made connections between the dotplots and histograms used to represent a distribution of simulated statistics and the density curves used to represent theoretical sampling distributions. At the most basic level, students noticed their visual similarities, with many using the term "Normal" to refer to discrete and continuous bell-shaped curves alike. When comparing simulated distributions and Normal curves, students often described one as an approximation of the other. In some cases, students' beliefs about the relationship between the two models shifted over the course of the year. Soon after theoretical sampling distributions were introduced, students often described them as an "estimate" or a "general overview" or even "imprecise" while simulated distributions were described as "exact results" or "the actual distribution." By the end of the year, students were more likely to believe that an infinite number of shuffles would result in the theoretical distribution. That is, students began to treat z, t, and χ^2 distributions as the true sampling distribution – an infinite, theoretical construct rather than a fallible model.

At a more advanced level, students made connections in terms of the outcomes being represented – for example, a distribution of statistics, or a "just by chance" distribution – or in terms of the distribution's purpose in the larger inferential process. In his individual interview, Ryan made rich connections between the two sampling distributions after using both approaches to test a difference of means. The data came from a randomized experiment that compared the effects of two treatments on blood pressure.

Ryan: Well with both of them you end up with a bell-shaped distribution to compare what you got initially in the first sample to. Yea. I mean they're – and both the distributions were centered at the difference being 0. So yea.

Catherine: Why are they centered at the difference being 0?

Ryan: Because the [t] distribution is shaped as if there is no difference … And then [in the simulated distribution], we just took that idea and put it into practice by just doing a bunch of samples of this assuming they had the same effect on blood pressure. And then we made our distribution out of samples instead of [pause] inference.

Catherine: Ok, so [earlier] you said these are both sampling distributions. Are they – would these be pictures of the same thing?

Ryan: No, because [the density curve] consists of t, the statistic of t, while [the simulated distribution] consists of the statistic of the mu difference (a difference of means). … Technically they're not the same, but they're both centered at 0, because a t distribution is kind of like a z distribution, where it measures the amount of standard deviations away. But a mu difference is just the actual difference between the two [means]. So it would be centered at 0. And this one makes sense that it's centered at 0, because if it's not different then it's no standard deviations away.

Ryan recognized that the *t* distribution represented a distribution of *t* statistics that could be obtained through repeated sampling; however, many students did not correctly identify the variable for a theoretical distribution. Although failure to identify the units in a randomization distribution is a concern in classes that employ simulation-based inference (Rossman & Chance, 2014; Saldanha & Thompson, 2002) the participants in this study tended to be more confident describing simulated distribution than they were describing theoretical distributions. Compare Natalia's descriptions of the two representations.

Natalia: (describing the Normal curve): That represents the distribution of our data, and if we did the experiment over and over again, it like represents what our results would look like. … The results of the test here when the child was asking being asked which thing it would choose.

Natalia: (describing the dotplot produced by the applet): Um, this dotplot is the distribution of the number of heads that we simulated when we were doing the coin flip. So in our case that would represent the proportion who chose helper.

When describing the Normal curve, Natalia seems to refer to a distribution of sample data rather a sampling distribution. Data like these confirm that the struggle to coordinate multiple levels, described in Chapter 4, is not unique to simulation-based inference.

Calculating P-values

When the applet counts the trials more extreme than the observed statistic, the resulting proportion can be interpreted as a p-value, and most students recognized the connection between theoretical and empirical p-values by the end of the year. Further, many students carried over the idea of counting samples when they described how the calculator found a p-value. For example, Grace described the calculation from both approaches as counting the results that fall beyond the sample data. Note she assumes the calculator is using the same statistic as the applet – the count of 14 successes in the original data – rather than a standardized statistic.

Grace: They’re similar in that you’re looking for a p-value and it’s going to measure the same thing … They’re both counting from 14 since that’s what we’re testing for, but then that’s just counting the number of repetitions whereas this is assuming that it’s Normal so this is the probability of it happening with a Normal distribution.

When asked how the calculator found a p-value, many students were able to draw a density curve and shade the area corresponding to the p-value; however, some students could not remember how a calculator found a p-value based on a test statistic.

These students often speculated that the calculator was doing the same thing as the applet.

Libby: This is just taking more possibilities into consideration. But [the traditional test] is one specific outcome whereas [the simulation-based test] is taking 100 different outcomes. Like they're both calculating p-values-ish, but that just seems like broader scale. Except for maybe in the traditional test, I just don't understand the magic of calculators and they're actually doing all of this and then we're just getting this – our t-value, our p-value, our degrees of freedom, all that.

Failing to remember the details of the traditional test, Libby made connections from simulation-based inference, which supported a useful conception of the p-value. In her interview, she repeatedly demonstrated a working conception of the p-value as the probability that the observed results occurred by chance, though she could not remember the details of the traditional test.

Although many students were unsure of how the calculator found p-values, they were nearly unanimous in the belief that the p-value given by the calculator was more accurate. They provided various reasons to support this belief. The quotes below come from a group interview with Ryan and Eva and individual interviews with Hannah and Hazel.

Ryan: I guess because… Well, I think this p-value is calculated a different way than this one. This one was calculated literally by counting how many of the shuffles were that out of 10000, when this one was just (pause) was um… was true (laughs) – was the real one. The real – the real thing.

Catherine: The real thing?

Ryan: Yea, I don't know how else to say it.

Eva: The real deal. It's calculated.

Hannah: Well we used [the calculator] most this year. And I felt like it was more accurate than doing a simulation, because you get a number that has like – like [the applet] just says 0, but [the calculator] also has numbers behind the zero, so it kind of feels more accurate.

Hazel: Maybe because the law of large numbers, [the p-values] get closer as [the number of trials] get higher, but we're not to infinity, so they're not the exact same.

Some students seemed to trust the p-value from the traditional test on the virtue of its being computed using a calculator, a trusted tool. Furthers, as Hannah pointed out, the calculator always produces a p-value to a large number of decimal places, which lends a (sometimes unwarranted) air of accuracy; in contrast, if a statistic is very unlikely, it may not appear at all in a set of simulated trials, resulting in an estimated p-value of 0. Lastly, beliefs like those expressed by Hazel were very common by the end of the year. Students correctly observed that empirically estimated p-values tend to be closer to theoretical p-values when the number of trials is large. However, it is important to note that both empirical and theoretical p-values are estimates; both empirical and theoretical models are imperfect representations of the underlying system.

Implications

Are These Connections Productive?

The results of this study suggest that students can make connections across traditional and simulation-based approaches, but do these connections lead to deeper understanding of traditional methods or the core logic of inference more generally? Because this was not an experimental study, there is no controlled comparison of pedagogical practices, but some proposed advantages of simulation-based inference seem to be substantiated in the data.

One oft-cited advantage of simulation-based inference is that it allows inference to be introduced early in the course and reinforced in various contexts, whereas traditional inference cannot be introduced without the machinery of theoretical sampling distributions (Holcomb, Chance, Rossman, Tietjen, et al., 2010; Tintle et al., 2011). On the day z-tests were first introduced in class, students filled out an exit ticket prompting them to make connections to the simulations they had used previously. On the day t-tests were introduced, students again filled out an exit ticket. Responses to these formative assessments suggest that students had already developed inferential conceptions by the point in the year when inference would be first introduced in a traditional course.

Hannah: z-test is similar to the 3S strategy, because they both have the same hypothesis and conclusion with p-value. They are also both Normal. Z-test is different in which it needed a formula; $z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$. Also it had three conditions it had to follow; random, normal and independent. Whereas, 3S didn't.

Isabella: z-test and 3S strategy are similar in that they both use the null hypothesis to check and see if the data happened by chance. Z-test and 3S strategy are different in that the 3S strategy does not test for Normal and Independent because the graph of the simulated data can show this.

Devon: A t-distribution is the bell-shaped distribution of statistics – the mean divided by the standard deviation of the sampling distribution. ($t = \frac{\bar{x}-\mu}{\frac{s_x}{\sqrt{n}}}$). This is similar to the problems that required z & t scores to find how far a value is from the mean & how often it occurs.

Having previously applied the logic of inference – students were accustomed to considering whether observed results could have occurred just by chance – they were able to focus on aspects of the tests that were new, such as the formulas for test statistics and the role of theoretical probability distributions.

Throughout the semester, students who provided the strongest conceptual descriptions of inference seemed to make connections to previous experiences with simulations. For example, in class periods two days apart, students responded to parallel formative assessment items. One asked students to describe the role of the simulation in SBI and the other asked students to describe the role of the z-distribution in a z-test. The additional prompts given also mirrored each other. Note the structural similarities as Ryan and Grace describe the two approaches.

Ryan: The simulation allowed us to see how often the Mythbuster's simulation would occur by chance. The simulated distribution represented the experiment done multiple times when yawning wasn't contagious. We make a decision about the hypothesis by seeing how likely to occur our result was.

Ryan: The z-distribution allows us to see how often a certain result would happen if the null were true. It represents how many st. devs away from the true value a number is. If the p-value we get from our number placed in a z-distribution is very low, it is unlikely our result would happen by chance.

Grace: The simulation was necessary b/c we needed to test if the results could happen "just by chance" or if they're statistically significant. We use the simulation distribution to show what differences happen just by chance and compare that to our data…

Grace: The z-distribution helps us gauge where our sample data lines up with the null (because the distribution assumes the null is true)…

The first exam that assessed traditional inference included a survey question asking students whether they thought experience with simulations had helped them understand z-tests and t-tests. Of the 10 students who responded, 7 said yes, and the most common reason given was appreciation for the visual representations provided by the applet. Students were also asked to share if and how simulations made hypothesis testing more confusing. In response, they mentioned the fact that each simulation is different, that having more options complicates the choice of test, and that these methods take a lot of time.

In response to questions about whether simulations had helped or caused confusion, several students provided answers without clear connections to simulations. For example, when asked if simulations had helped, Alicia said, "No, because I am still not sure what the differences between a z-test and a t-test are. I mean logically it doesn't make sense to me." This difficulty is certainly valid, but it is more closely related to traditional tests than simulation-based tests. Responses like this suggest that students do not always clearly distinguish the two approaches.

On that early survey, Eva responded that simulations made it difficult for her to "grasp the concept" of inference. After encountering both approaches in class, participants in this study widely agreed that traditional tests were easier than simulation-based tests. For example, the data for this study included 20 task-based interviews, and in each, students were given a choice of which approach to use first; only one student completed the simulation-based test first. This may not be surprising, given that participants in this study had considerably more experience with traditional methods. Although students found simulation difficult, several students volunteered statements about the value of simulation-based methods in their final interviews. The quote below illustrates how Eva's position changed.

Eva: But like when you just do [the traditional test], and you're like, "Okay, p-value." I just rely on "p-value is less than this so reject the null." Like I don't actually think about it.

Catherine: Mmhmm, you just sort of follow the rule.

Eva: Yea, which I mean is a good rule for passing the AP test, but for actually thinking about statistics, it's like [the traditional test] doesn't really mean anything to me. And like making a conclusion, it's just, I don't know. [In the simulation approach], I'm like, "Okay, I understand why," and [in the traditional approach], I'm just like, "Rules!" And write it.

Quotes like these raise important questions for both teachers and researchers. What aspects of simulation-based inference lead to productive struggle for students? These students had considerably more experience with traditional methods. Did simulation provide a challenge simply because it disrupted students' problem-solving routines or is it beneficial in other ways? Further, traditional inference methods are not a monolithic pedagogical approach. How can teachers of traditional courses discourage over-reliance on rules and procedures?

Hyperconnections

In general, students are more comfortable describing the similarities between approaches than they are describing the differences. Several examples of these hyperconnections between approaches have already been mentioned. Some students are overly liberal with the term “Normal”. Some assume that all p-values are calculated by counting simulated trials. Some fail to recognize the role of standardized test statistics in traditional inference. The goals of the particular course may dictate the level of concern about these kinds of errors. Some teachers may be concerned by the failure to distinguish approaches yet pleased with connections between conceptually similar procedures.

However, other hyperconnections between approaches are unambiguously problematic. While working on inference tasks, students commonly combined the two procedures, usually incorporating traditional procedures into a simulation-based approach. For example, after using their physical model to simulate one trial, students often tried to use the results of the simulation as basis for a traditional test. For example, on a chapter test, Isabella designed a simulation to test whether seagulls have a preference for where they land. In the study, the outdoor space was made up of 56% sand, 29% mud, and 15% rocks, and biologists recorded the landings of 200 seagulls.

Isabella: First you would use a spinner and label 56% of it “sand”, 29% of it “mud”, and 15% of it “rocks”. Then we would spin the spinner 200 times. We would then count up how many of these spins landed on sand, mud, or rocks and organize them into 3 columns. To get the expected counts, we would multiple 56%, 29%, and 15% to 200 for sand, mud, and rocks. Then

> we would enter the number of spins for sand, mud, and rocks we collected from the simulation and the expected counts into 2 different lists on a calculator. We would then calculate χ^2, p-value, and df using the calculator function $\chi^2 - \mathrm{GOF}$.

Isabella's model is appropriate for simulating where seagulls would land if they had no preference. However, instead of repeating the process many times, she used one simulated trial as the basis for her conclusion. In Chapter 4, errors like this are described as difficulty transitioning from the level of sample data to the level of the sampling distribution.

Less commonly, students simulated many trials, but combined approaches as they calculated a p-value. For example, a few students simulated many trials to create a distribution of statistics, but started counting from the point of the z statistic or even the p-value obtained from the traditional test rather than the statistic recorded in the simulation. This error may only occur in situations where students are asked to use both approaches consecutively.

These hyperconnections serve as a reminder that some of the proposed advantages of simulation-based inference – e.g., avoiding mathematical formulas and theoretical sampling distributions – do not apply to courses that use simulations in addition to traditional tests. As described, using two sets of inferential methods in instruction can lead to confusion as students make inappropriate connections between the traditional and simulation-based approaches. When deciding how to teach inference, teachers must weigh the benefits of productive connections against the costs of inappropriate connections.

Recommendations

Over the course of several years, modifications have been made to the course taught at PKY in order to capitalize on potential connections between traditional and simulation-based methods. For teachers who choose to use simulations as complements to theory-based methods in instruction, the following recommendations may support productive connections between the two.

Engage Students in Simulation

The course description above mentions three ways to engage students in simulations: using physical chance devices to carry out simulation, using technology to carry out simulations, and designing simulations. Each may be beneficial, and there are a number of considerations for incorporating these modes of engagement strategically.

Many (e.g., Chance et al., 2004; Holcomb, Chance, Rossman, Tietjen, et al., 2010) have recommended starting with hands-on physical simulations before transitioning to computer simulations within the same context. Chance et al. (2004, p. 315) suggest that physical simulations “give [students] a meaningful context to which they can relate the computer simulations. Otherwise the computer provides a different level of abstraction and students fail to connect the processes.” Mills (2002) reports a general consensus in the literature that physical and computer simulations complement each other; however, few of the articles she reviewed were empirical studies. Teachers may choose physical simulations early in the year when students are less familiar with the process of randomization.

Although physical simulations are often recommended as a starting point, empirical p-values are not reliable estimates of the true likelihood unless the number of trials is large, which is more feasible using computer simulations. Technological tools allow students to carry out the complete 3S process on their own in a relatively short period of time. Applets are free and relatively simple to use; these are particularly appealing advantages for courses in which simulation is not the primary means of conducting inference. Two popular sets of web applets are *StatKey* (http://lock5stat.com/statkey/), which is associated with the authors of *Unlocking the Power of Data,* and *Rossman/Chance Applets* (http://www.rossmanchance.com/applets/)*,* which is associated with two authors of the *Introduction to Statistical Investigations Curriculum.* Other technological tools, such as Tinkerplots (C Konold & Miller, 2011), offer more flexible modeling capabilities but require a larger investment of class time to master the software and, in some cases, money for the software licensing fee.

After participating in teacher-designed simulation activities, students may benefit from designing their own simulations. As discussed in Chapter 3, choosing and evaluating a design often leads student to consider the role of randomness and the null hypothesis. Further, student-designed models may be thought-revealing, allowing the teacher access to inferential conceptions not visible through more routine inference procedures.

Be Intentional about Transitions between Approaches

The course description above illustrates one way to transition from simulation-based inference to traditional inference. At the 2015 United States Conference on Teaching Statistics, three authors of the textbook *Statistics: Unlocking the Power of Data* proposed another. After first conducting a simulation-based test, they find the p-value using a Normal approximation, where the standard error is the standard deviation from the simulated distribution. Next they transition to a z statistic, still using the standard error from the simulation. Finally, they introduce a formula for the standard error, demonstrating to students that all four approaches result in similar p-values.

Whether you start by transitioning to a standardized z-statistic or to a Normal approximation, be intentional about the transition from a simulation-based approach to a traditional approach. Similarly, if you introduce formal language and inscriptions alongside traditional tests, be explicit about the connections to informal language. For example, students may not immediately realize that the null hypothesis is a formal name for the "just by chance" explanation. Providing students with a framework to make connections capitalizes on the inferential conceptions students have already developed. Further, formative assessments that prompt students to reflect on these connections may provide benefits to both teachers and students.

Don't Neglect Traditional Representations

In classes that use inference-capable calculators to find p-values, students may not spontaneously produce graphical representations of the theoretical distribution. However, absent those representations, students may come to regard the calculator as a mysterious black box. Perhaps counterintuitively, lack of familiarity with the mechanics of p-value calculation is associated with a tendency to overrate the accuracy of p-values estimated using traditional methods. If the details of traditional inference are important goals of the class, it is worthwhile to continue using traditional representations – such as formulas for standardized test statistics and shaded density curves to represent p-values – at least intermittently throughout the course.

Emphasize Modeling in Both Approaches

Both theoretical probability distribution and simulated sampling distributions are *models* for the outcomes that would occur just by chance under the null hypothesis.

However, students tend to view one as the true distribution and the other as an approximation; by the end of the year, most students in this study put more faith in the traditional approach. Thus, students may be very surprised to find that the p-values obtained from the two approaches are not exactly the same even when the number of trials is very large.

A simulation-based approach offers many opportunities to discuss modeling, as a physical model is chosen to match the data scenario and the randomness inherent in the study design. However when using the traditional approach, students do not always recognize checking conditions as a way to evaluate the fit of a theoretical model. One issue is that the complicated guidelines for checking conditions can conceal the commonality across all traditional tests: namely, theoretical models only fit well when the sample size is reasonably large. When asked whether she checked the Normal condition for a chi-square test, Isabella gave the following response.

Isabella: No. Because the χ^2 statistic doesn't depend on how small or large a sample or group size is. Instead it depends on if the expected counts are larger than 5.

The expected counts will only be larger than five when the sample size is reasonably large, but the rule of thumb, expressed in terms of expected counts, obscured the connection to sample size. To avoid this confusion, refer to the "sample size condition" for all traditional tests.

Perhaps related to broader misunderstandings about the conditions for inference, students often believe that the "Normal condition" applies to simulation-based tests and traditional tests alike. One way to address this confusion is to use a simulation-based test as an alternative when the conditions for a traditional test are not met. It is problematic to require students to check conditions but only provide data where the conditions are met. Students may come to believe that the conditions are only a formality and that the theoretical distribution can be regarded as “true”, when in fact, the simulation-based model is sometimes a more appropriate model for the data.

Summary

This article described an AP Statistics course that employed both traditional and simulation-based inference methods and aimed to emphasize the connections between the two. As illustrated, students enrolled in the course often made productive connections across approaches, which provides preliminary support for the idea that complementing theory-based methods with simulations can deepen conceptual understanding of inference. However, students were also prone to “hyperconnections” – overgeneralizing the characteristics of one approach to the other, or even combining traditional and simulation-based procedures. The article offered four recommendations to capitalize on potential connections between traditional and simulation-based methods in courses that employ both to introduce the logic of inference. However, considerable work remains for teachers and researchers as novel pedagogical approaches for teaching inference are implemented and refined.

CHAPTER 6
CONCLUSIONS AND DISCUSSION

Given its pervasiveness in the scientific research literature (Nickerson, 2000), rich understanding of statistical inference is an important goal for students in introductory statistics classes – both those who intend to produce their own statistical analyses and those who will engage with statistics primarily as critical consumers of data-based reports (ASA, 2016). Although significance testing is a ubiquitous data analysis tool, there is evidence to suggest it is misunderstood by many who use it (Nickerson, 2000). Thus, statistics educators have devoted considerable effort to reforming inferential instruction in recent years; in particular, many have proposed simulation-based inference methods as a means to improve understanding of inference (e.g. Cobb, 2007; Garfield et al., 2012; Lock et al., 2014; Tintle et al., 2011). As enrollments in statistics courses grow and simulation-based inference methods gain popularity (ASA, 2016), a research-based understanding of the impact of simulation-based inference becomes necessary.

This dissertation examined how students used traditional and simulation-based inference methods to understand inference in the context of an AP Statistics course. Using a modified grounded theory methodological approach (Charmaz, 2014), the conclusions of this study were drawn from systematic, inductive analysis of data collected in an AP Statistics course. Data collection and analysis were informed by a models and modeling perspective (Lesh & Doerr, 2003a), which assumes that reality is accessed through models and representations that emphasize different aspects of the underlying system. This theoretical perspective supported the study's focus on representations, like mathematical equations, graphs, concrete materials, and notation systems (Seel, 2014) as mediators of inferential understanding.

The study was situated in the context of an AP Statistics course taught by the author at P.K. Yonge Developmental Research School (PKY). The AP Statistics curriculum prescribes traditional inference methods and includes nine tests of significance (College Board, 2010). In total, statistical inference – which includes tests of significance and confidence interval estimation – constitutes 30-40% of an AP Statistics course (College Board, 2010). In addition to the prescribed AP Statistics curriculum, the course taught at PKY regularly incorporated simulation-based inference methods; thus all students in the study were exposed to both inferential approaches. The course incorporated simulation-based inference throughout the year and added traditional inference later in the second semester. In total, the course included fourteen in-class experiences with simulation-based inference, including multiple opportunities for groups of students to design their own simulations and carry out simulations using physical chance devices and applets. However, because the AP Statistics course description emphasizes traditional inference, students had considerably more experience with traditional, theory-based methods by the end of the school year.

Data were collected in two phases. First, a pilot study was conducted at the end of the 2013-2014 academic year. In this first phase, individual task-based interviews (Maher & Sigley, 2014) were conducted with seven students, selected to represent a range of statistical achievement in the class that year. In the second phase, data were collected from all eleven students enrolled in the course in the 2015-2016 academic year. Data collected in the second phase included student responses to targeted formative assessments, exam items, and survey items; daily field notes and journal entries written by the teacher-researcher; and transcripts of individual and group interviews. Because grounded theory encourages simultaneous data collection and analysis, data collection was informed by the analysis, as the researcher aimed to saturate emerging categories.

Data analysis entailed a process of systematic coding in multiple phases, according to the guidelines for grounded theory presented by Charmaz (2014); these guidelines provided a systematic yet flexible way to study the emerging data. In the initial coding phase, each segment of data was assigned a concrete and descriptive code intended to reflect the students' actions. After comparing these initial codes to the data and looking for patterns in the codes across interviews, the researcher inductively constructed a set of focused codes to "sift, sort, synthesize, and analyze" the large amounts of data (Charmaz, 2014, p. 138). The codes reflect the "sensitizing concepts" of the models and modeling perspective, but these concepts were not accepted into the analysis until they could be substantiated in the data (Charmaz, 2014; Corbin & Strauss, 1990). Throughout the study, memos were used to document the process of coding and to draft descriptions of conceptual categories. Ultimately, the focused codes were grouped into categories, which were presented in three articles in Chapters 3, 4, and 5. This chapter synthesizes the findings of those articles, interpreting the results in relation to the overarching research question. Limitations, implications, and directions for future research are also discussed.

Synthesis of Results

The central research question of this study asked how students use traditional and simulation-based inference models to understand inference. The data suggest that the participants interacted with the models, representations, and tools of the two approaches in qualitatively different ways. Using the tools made available in this course – including a four-step problem-solving framework, a memorized list of conditions for inference, and a graphing calculator with wizards for statistical tests – many students were able to carry out a traditional significance test quickly and efficiently by end the of the year. However, when students were probed for details about the underpinnings of the traditional, theory-based approach in individual interviews, some revealed incomplete knowledge of the computational mechanics or perhaps more concerning, the underlying logic of inference. Further, traditional inference tasks, similar to tasks these students had completed in preparation for the AP Statistics exam, rarely led to spontaneous discussion of statistical modeling or the logic of inference. Thus, these tasks did not provide many opportunities for students to challenge each other's conceptions or for a teacher/researcher to evaluate student thinking.

Additionally, students' use of tools and representations for traditional inference sometimes differed from the expectations of the teacher-researcher. For example, graphing calculators served multiple purposes, not only as a data analysis tool, but also a memory aid to help students choose a significance test or revise their hypotheses. On the other hand, when not required to draw a density curve as a representation of the theoretical sampling distribution, graphical representations were used less than expected for traditional inference. Individual interviews suggest that some students do not recognize the referent of the density curve graph, which may explain why they do not routinely employ them in their reasoning.

In contrast to the relative ease with which they carried out traditional inference, many students struggled to design a simulation and use the simulation-based model to draw conclusions about statistical significance. At the design stage, students sometimes proposed and evaluated multiple models before settling on one suitable for a “just by chance” explanation of the data. As evident from discussions among students, these modeling decisions often prompted consideration of the null hypothesis and the source of randomness in the study design. In some cases, student-designed models revealed student thinking that had been obscured in their more successful use of traditional methods.

After choosing an appropriate physical model, some students still struggled to carry out simulation-based inference. The common errors that arose at this stage can be described largely in terms of two challenges. First, students struggled to coordinate the multi-level scheme, which includes the population or true underlying relationship, the distribution of single sample, and the distribution of statistics collected from multiple samples. For example, some students confused the sample size of a single sample with the number of samples in an empirical distribution of statistics. Others did not easily transition from the sample level to the sampling distribution level, which sometimes led to inappropriate interpretations of a single simulated sample. Second, students struggled to coordinate two perspectives: the real-world where the sample data was collected and the hypothetical perspective where the null hypothesis was assumed to be true. This led some students to imagine a distribution produced through simulation based on the null hypothesis as a distribution produced through replication in the real-world. This conception was manifested as a number of different errors. Despite these difficulties, finding an empirical p-value using the applet as a tool prompted many students to consider the p-value as a probability of observed data and a tail probability.

Students often demonstrated incomplete understanding of inference through their use of simulation-based methods, but the findings do not imply that these methods caused confusion; all participants in the study were exposed to both traditional and simulation-based methods in instruction, so no conclusions can be drawn about the causes of their conceptions. However, there is some evidence to suggest that students exposed to both traditional and simulation-based methods in instruction can make productive connections across approaches. For instance, students who find theoretical probability models and their representations opaque may describe statistical concepts in terms of more accessible empirical models. Although students consistently chose traditional methods when offered the choice between approaches, several students stated that they found simulation-based approaches useful for understanding *why* inference methods work.

The data in this study include numerous examples of productive connections between traditional and simulation-based approaches. However, students tend to be more comfortable describing similarities between approaches than they are describing differences. Students sometimes overgeneralize, attributing characteristics of one approach to another. Even more problematic, students sometimes combine the traditional and simulation-based approaches as they carry out a single inference task. These inappropriate "hyperconnections" serve as a reminder that some proposed advantages of simulation-based inference – e.g., avoiding mathematical formulas and theoretical sampling distributions – do not apply to courses that use simulations in addition to traditional tests; instead, simulations add more models and representations for students to consider. When deciding which inference methods to include in instruction, teachers must weigh the benefits of productive connections against the costs of inappropriate connections.

Prior to this study, the statistics education literature provided little description of how students employ the tools and representations of traditional and simulation-based inference models. In particular, little had been written about the pedagogical approach of complementing traditional inference with simulation-based methods or the use of this approach in a high school setting. Thus, the findings of this study represent a contribution to the research literature in statistics education.

Interpretation of Results

In contrast to the statistical generalizability of quantitative studies, this qualitative study aims for *analytical generalizability,* which involves "a reasoned judgment [on the part of the reader] about the extent to which the findings from one study can be used as a guide to what might occur in another situation" (Kvale, 1996, p. 233). Even in the field of statistics education, which has traditionally favored quantitative methodology (Gordon, Reid, & Petocz, 2010), there have been calls for qualitative research that produces "vivid descriptions that … represent a researcher's well-formulated perspective rather than an objective reality that cuts across a tightly specified range of context" (Groth, 2010). Accordingly, this dissertation describes how students used traditional and simulation-based inference to understand inference in the context of an AP Statistics class; the study's conclusions are supported by data co-constructed by the teacher-researcher and participating students. This section provides recommendations for how the study’s results might be interpreted in light of the study’s design and epistemological position.

First, all data were collected in AP Statistics classes taught by a single teacher-researcher at a single school. It is not reasonable to assume that the results of the study will generalize to all introductory statistics courses that complement traditional inference with simulation-based inference. Factors including the teacher, the textbook, the school environment, and individual student traits likely affected how the participants in this study used traditional and simulation-based inference models to understand inference. In order to help the reader “discover the extent to which the theory does apply and where it has to be qualified for the new situations” (Corbin & Strauss, 1990, p. 15), this dissertation includes thorough descriptions of the context, participants, and the inferential instruction used in class.

Second, the study provides a naturalistic description of the use of models, tools, and representations in a single classroom environment; it is not an experimental design. Thus, the study does not provide a basis for comparing the effectiveness of traditional and simulation-based inference, as all students were exposed to both approaches in instruction. Further, the study does not provide a basis for comparing simulation-based inference with other possible uses of class time, such as additional practice with traditional methods. However, this study may nevertheless inform the choice of curriculum by providing detailed descriptions that illuminate the reasoning processes each approach elicits.

Lastly, the research process in this study is subjective. In contrast to the positivist leanings of early grounded theorists, this study acknowledges the subjectivity of both data collection and analysis. Recognizing that data are not merely collected but co-constructed by the researcher and the participants, this dissertation described how the data were elicited with attention to the role of the researcher and the influence of concurrent analysis. Specifically, three facets of the study's subjectivity merit mention.

First, the researcher's dual role as teacher necessarily influenced the study. The teacher-researcher chose to use simulation-based inference in her class, which suggests a certain pre-existing belief in the potential of these methods. Second, the researcher was familiar with views of inference shared among statisticians and statistics educators before she began this study. This lens of disciplinary knowledge likely influenced the researcher's perceptions of the students' modeling practices. Third, incidents in the data were coded by a single person, so there is no indication of inter-rater reliability. This highlights the importance of the teacher-researcher's perspective in constructing the final analysis. Thorough documentation of data analysis and clear, detailed examples "invite the reader to appraise [the researcher's] interpretations and think about other ways the data could have been interpreted" (Kalinowski et al., 2010, p. 30).

Implications and Future Research

The findings of this study have implications for teachers and researchers interested in the teaching and learning of statistical inference. At the same time, the study raises further questions to be answered in practice and through research.

First, this study provides a qualitative description of the use of traditional and simulation-based inference as complements in instruction. This description may be of use to teachers as they consider which inference methods to include in their courses, though as mentioned above, this study does not constitute a comparison of approaches.

In addition to existing quantitative comparisons of student achievement across curricula, there is a need for qualitative comparison of the two approaches. However, this study suggests that such comparisons will present a methodological challenge. In particular, how will future studies account for the different ways in which students interact with the two models? Future work should acknowledge that a study pitting pedagogical approaches against each other necessarily imposes assumptions about the learning processes and outcomes that are valued. Further, it is not accurate to talk about traditional, simulation-based, or complementary approaches as if they are monolithic pedagogical styles that can be applied uniformly in any classroom context. Future work should attend to context, as students' understanding of inference is likely influenced by many factors working in concert.

Second, this study highlighted the complex relationship between student reasoning and tools for inference: Not only do students use tools, they interact with them. For example, a tool may serve as a memory aid or as a source of cognitive dissonance; it may prompt students to discuss certain statistical concepts while enabling them to disregard others. These findings have implications for teachers, as they decide which tools should be made available to students. In addition to other factors like availability and ease of use, the impact on students' reasoning is a nontrivial consideration in choosing a tool, and there is need for future work that investigates the impact of specific tools on student reasoning.

Third, this study identifies errors that commonly arise in courses that employ simulation-based inference and characterizes the conceptions underlying those errors. Awareness of these issues may have implications for instruction, both in proactive instructional design and reactive responses to students. However, there is need for future work to explore how to address these errors effectively.

Lastly, whether used alone or as a complement to traditional inference, simulation-based methods represent a substantive disruption of the status quo in introductory statistics courses. Thus, most teachers who undertake these methods – by choice, because of their inclusion in the standards, or because of an administrative decision – will be implementing methods dramatically different from those they were taught. Further, as illustrated throughout this report, simulation-based methods do not eliminate the difficulties of statistical inference. Thus, there is need for work by statisticians and teacher educators to support teachers as they implement new inference approaches and meet the challenges that subsequently arise.

In summary, this dissertation provided a qualitative description that informs our understanding of how students use traditional and simulation-based inference models to reason about inference in a class that employs both in instruction. In particular, the study identified common errors that arise in simulation-based inference and discussed the connections that students make between the two models and representational systems. The data collected substantiate the educative potential of simulation-based inference methods. However, use of simulation-based methods – alone or as a complement to traditional inference – is not a panacea. Considerable work remains for teachers and researchers as various pedagogical approaches are implemented and refined in classrooms.

APPENDIX A
DATA COLLECTION TIMELINE

Table A-1. Data collection timeline.

Academic year 2013-2014	Teaching of AP Statistics using both traditional and simulation-based methods to introduce the core logic of inference
May 2014	Pilot study: Task-based interviews with 7 AP Statistics students
Academic years 2014-2016	Modifications to instruction based on the results of the pilot study
Spring 2016	Targeted formative assessments collected about once per week Five chapter tests that included assessments of simulation-based inference Daily journal entries On-going analysis of data with memos to document the process of coding and theory development
May 2016	Task-based interviews using piloted interview protocol

APPENDIX B
INFORMED CONSENT FORM

Informed Consent

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

Purpose of the Research Study:
The purpose of this study is to describe how students reason about statistical inference. More specifically, we have used two kinds of inference methods in class this year: traditional methods (e.g., z-tests and t-tests) and methods based on simulations. This study explores how students use traditional and simulation-based inference methods to understand the logic of inference and what connections students see between the two methods.

What students will be asked to do in the study:
Near the end of the school year, students will be asked to participate in a group interview and an individual interview. During the interviews, students will be asked to "think aloud" as they use traditional and simulation-based inference methods. Students will also be prompted to compare and contrast the approaches, describing the connections perceived between the two.

Time Required:
Each interview will take 30-45 minutes to complete. The group interview will be conducted during class time, and the individual interview will be conducted outside of class.

Risks and Benefits:
There are no risks associated with participating in this study.

Compensation:
There is no compensation for participating in the study.

Confidentiality:
Student identity will be kept confidential to the extent provided by law. Student information will be assigned a code number or pseudonym. Student names will not be used in any report.

Voluntary Participation:
Participation in this study is completely voluntary. There is no penalty for not participating.

Right to withdraw from the study:
Students have the right to withdraw from the study at any time without consequence.

Whom to contact if you have questions about the study:
Catherine Case, Doctoral Candidate, School of Teaching and Learning, University of Florida phone: 256-454-5348; e-mail: ccase@ufl.edu

Whom to contact about your rights as a research participant in the study:
UF IRB Office, PO Box 112250, University of Florida, Gainesville, FL 32611
phone: 352-392-0433; email: irb2@ufl.edu

Signatures: (*Please place an X on the appropriate lines.)*

____ I have read the procedure described above. I voluntarily **give my consent** for my child, ___________________________________, to participate in the study. I have received a copy of this description.

____ I have read the procedure described above. I **do not give my consent** for my child, ___________________________________, to participate in the study. I have received a copy of this description.

Parent/Guardian: ____________________________________ Date: ________________

____ I have read the procedure described above. I voluntarily **give my assent** to participate in the study. I have received a copy of this description.

____ I have read the procedure described above. I **do not give my assent** to participate in the study. I have received a copy of this description.

Student: ____ Date: ________________

APPENDIX C
INTERVIEW PROTOCOLS AND TASKS

Individual Interview Protocol

[Let the student choose inference technique—traditional or simulation-based inference.]

- Why did you choose do use this approach (traditional/simulation-based) first?
- While you're carrying out the test, think out loud so I can understand what you're doing.
  - What does this part mean?
  - How did you know to do that?
  - Can you draw a picture to show me what you mean?
- What did you conclude about the results of the study?
- Write down an interpretation of the p-value in the context of this problem. What is it the probability of?

[Ask the student to work through the problem again using the alternative inference approach.]
If the student is unable to complete the task provide assistance. If they struggle with traditional inference, remind them of the State, Plan, Do, Conclude framework. If they struggle with simulation-based inference, guide them with prompts like the following:

- What claim are the researchers trying to find evidence against? What are they trying to find evidence for?
- Which of these two claims should we try to model?
- How can we model this claim using chance devices like coins, dice, or cards?
- What would you expect to happen if we continued with this process?
- Let's use a computer applet to make this process faster.
- How do we use the information given by the applet to draw a conclusion?

After the task has been completed using the two different approaches, prompt the student to make connections between the two.

- How are the two approaches you used similar?
- How are the two approaches you used different?
- After they've compared and contrasted the two on their own, direct their attention to specific parts of their work and ask if they see further connections between the two approaches.
  - Claims/hypotheses
  - Conditions
  - Calculations needed to find the p-value
  - Representations of the sampling distribution
  - P-values
- Are there any revisions you'd like to make to your p-value interpretation?
- Why do you reject the null when the p-value is small?

Group Interview Protocol

Instructions:

- Make students aware of the available tools (formula sheet, graphing calculator, chance devices, and computer).
- Introduce the inference task.
  - Does this result provide convincing evidence of response bias? Work together to decide.
  - Let the student choose an inference technique—traditional inference or simulation-based inference.
- If students ask what to write down: “Paper is here so you can show each other what you’re thinking. You don’t have to write anything down for me.”

After completing the first inference approach:

- So you used a [traditional/simulation] approach to decide if these results provided convincing evidence of response bias. Can you show me a different approach using a [traditional test / simulation]?

After completing both inference approaches:

- Now you have used two different inference approaches to analyze this data. Were the outcomes of the two tests consistent with each other? Is this surprising?

Prompts specific to simulation-based inference:

- Before you use the computer, can you show me how to use a physical simulation in this context?
- The applet shows a simulation using cards. Is this equivalent to the one you designed?

Interview Tasks

Task 1: Helper vs. Hinderer

In a study reported in *Nature*, researchers investigated whether infants take into account an individual’s actions towards others in evaluating that individual as appealing or aversive, perhaps laying the foundation for social interaction. In one component of the study, 10-month-old infants were shown a “climber” character that could not make it up a hill in two tries. Then they were shown two scenarios for the climber’s next try, one where the climber was pushed to the top of the hill by another character (“helper”) and one where the climber was pushed back down the hill by another character (“hinderer”). The infant was alternatively shown these two scenarios several times. Then the child was presented with the two characters from the video (the helper and the hinderer) and asked to pick one to play with. The researchers found that 14 of the 16 infants chose the helper over the hinderer.

Task 2: Oil and Blood Pressure

In a study reported in the *New England Journal of Medicine*, researchers investigated whether fish oil can help reduce blood pressure. 14 males with high blood pressure were recruited and randomly assigned to one of two treatments. The first treatment was a four-week diet that included fish oil, and the second was a four-week diet that included regular oil. At the end of the four weeks, each volunteer’s blood pressure was measured again and the reduction in diastolic blood pressure was recorded. The results of this study are shown below. Note that a negative value means that the subject blood pressure increased.

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

Task 3: Response Bias

In Chapter 4 we learned how characteristics of an interviewer can lead to response bias. Two AP Statistics students decided to investigate this issue. They speculated that students would be more likely to identify as feminists if asked by a female interviewer. A sample of 60 male high school students were asked, “Are you a feminist?” Half were randomly assigned to a male interviewer and half were randomly assigned to a female interviewer. Of the 30 asked by a male interviewer, 11 responded, “Yes.” Of the 30 asked by a female interviewer, 15 responded, “Yes.

APPENDIX D
TARGETED ASSESSMENTS

Targeted Assessments (by date)

Research Questions

1. How do students use traditional inference models and simulation-based inference models to understand inference?
2. What conceptions of inferential topics do students hold, and how are these related to commonly occurring student errors?
3. What connections do students see between the two models and representational systems?

Date	Assessment (brief version)	RQ
1/15	Individual exit ticket • In your own words, what is a sampling distribution?	2
1/22	Individual exit ticket • In your own words, what is a sampling distribution? • Today we used an applet to simulate a distribution of means. How is that simulation related to the Normal distribution?	2 3
1/29	Ch. 7 exam, item 1 – Interpreting German tank simulation • Label the three parts of the model • In the dotplot in part 2, there is a dot at 149. What does this dot represent? In the dotplot in part 3, there is a dot at 151.5. What does this dot represent? • Is $\widehat{N} = Q_3 + 1.5 * IQR$ an ideal estimator for the number of German tanks in the population? Why or why not?	2 2 2
1/29	Ch. 7 exam, formative assessment (ungraded) • In class, we used an applet to simulate repeated sampling from a candy machine with 50% orange candies. Explain what the dotplot below represents. • Your textbook uses the image below to represent the sampling distribution of $\hat{p}$. Compare and contrast these two graphical representations – one a dotplot and one a smooth curve. How are they similar? How are they different?	2 3

2/12	Individual exit ticket • Explain how a z-test is similar to the 3S strategy we've used before. • Explain how a z-test is different from the 3S strategy.	3 3
2/17	Group activity • Your class notes show computer output for a z-test. Now use the 3S Strategy to test the same hypotheses. Add notes that explain how this process compares and contrasts to a z-test.	1 3
2/19	Individual exit ticket • In your own words, what is a t distribution? • Is it related to anything we've already learned?	2 3
2/26	Ch. 9 exam, survey (ungraded) • On Item 3, you had the option to use simulation-based inference or traditional inference to test a claim. How did you decide which one you wanted to use? • All year, we've been using simulations to test hypotheses. In this chapter, we learned about two traditional tests – a z-test and t-test. Do you think experience with simulations helped you understand z-tests and t-tests. If so, how? • Do simulations make hypothesis testing more confusing in some ways? If so, how? Be as specific as you can.	1 3 3
2/29	Individual in-class assessment • Based on their experiment, the Mythbusters confirmed that yawning is contagious. 1 minute survey: "I agree with their conclusion, because…" or "I'm skeptical about their conclusion, because…"	2
2/29	Individual exit ticket • Today we did a hypothesis test about the difference between two proportions using the 3S Strategy. Write a few sentences about the role of the simulation. • The following questions may help you get started: Why is it necessary? What does the simulated distribution represent? How do we use the simulation to make a decision about the hypotheses?	1, 2, 3

3/2	Individual exit ticket	
	• Today we did a hypothesis test about the difference between two proportions using a z-test. Write a few sentences about the role of the sampling distribution (the z distribution). • The following questions may help you get started: Why is it necessary? What does the z-distribution represent? How do we use the simulation to make a decision about the hypotheses?	1, 2, 3
3/12	Ch. 10 exam, item 3	
	• Suppose you want to use simulation-based inference to answer this question. Describe the simulation you would use to estimate the sampling distribution. Be sure to mention the device (e.g. coins, dice, cards), how you would identify outcomes (labels), and what variable or statistic you would record each time.	1
	• Suppose you want to use a t-test to answer this question. Describe the theoretical sampling distribution you would use. Be sure to mention the shape, center, and spread[1] of the distribution.	3
	• The p-value for this study is 0.064. A p-value is a probability. What is it the probability of? Explain in context.	2
3/14	Group activity	
	• Design a simulation to find out what values of χ^2 are likely to occur by chance. • Your design can use anything you want: spinners, bags of beads, dice, coins, cards, calculators, …	1

[1] This question is flawed, because we had not learned the standard deviation of the t distribution.

4/1	Ch. 11 exam, item 2, part 3 • Did you check the Normal condition before performing this test? Explain why or why not. Ch. 11 exam, item 4 • Suppose you want to use simulation-based inference to decide whether seagulls show a preference for where they land. Describe the simulation you would use to estimate the sampling distribution of the χ^2 statistic. (Your design can use any device you choose: spinners, beads, dice, coins, cards, calculators,…) • Describe the distribution of statistics that would be generated by your model. What would it look like?	2, 3 1 2, 3
4/4	Individual exit ticket As a class, we designed a simulation to test the association between row number and test score. I summarized student ideas on the board and numbered the steps. • For each step, write one or two sentences explaining why. (instructions given verbally)	1, 2
4/15	Ch. 12 exam, item 1 Context: Students are given a scatterplot and a regression line summarizing the effect of sugar on the life of cut flowers. A description of a simulation and 12 simulated slopes are also provided. • The average of the simulated slopes is near 0. Is that surprising? • Use the results of their simulation to estimate the p-value. Do you think this is a good estimate? Why or why not? • Interpret the p-value as a probability. (What is it the probability of?) Make sure your answer is in context.	2 2 2

APPENDIX E
SIMULATION-BASED INFERENCE ACTIVITIES

Date	Analogous Test	Activity (Source)	Context	Physical Simulation	Applet or other technology
8/24	Exact binomial test	Coke vs. Pepsi (adapted from this article by Floyd Bullard)	Each student tastes two unlabeled cups of soda and writes down whether think the cups contained Coke or Pepsi. The number of correct guesses is recorded. Do the results of the class taste-test provide convincing evidence that the students weren't just guessing?	Chance device: coins • Heads – correct identification of soda • Tails – incorrect identification • Flip coin once for each student in the original taste-test. • Record the number of correct guesses.	Applet: One proportion inference
9/4	t-test for difference of independent means	Sleep deprivation (Holcomb, Chance, Rossman, Tietjen, et al., 2010)	21 subjects were randomly assigned into two groups: $n_1 = 10$ in the unrestricted sleep group, $n_2 = 11$ in the sleep deprivation group. Data provided show the subjects' improvements between pre-test and post-test on a test of visual discrimination? Do the data provide convincing evidence that sleep deprivation caused the scores to be lower?	Chance device: cards • Write improvement scores on blank cards. • Put cards from both groups together and shuffle. • Deal cards into two piles to mimic random assignment. • Record the difference in means for the two groups.	Applet: Randomization test for two means

9/9	t-test for difference of independent means	Memorizing letters (Zieffler & Catalysts for Change, 2013)	Each student is randomly assigned to receive one of two lists of letters. Students are given 30 seconds to memorize as many letters as they can; when time is up, the students lists as many letters they can from memory. The number of correct letters in the written list *before* the student made a mistake is recorded. Do the results of the class experiment provide convincing evidence that one list is easier to memorize than the other?	Chance device: cards • Write number of letters correct on blank cards. • Put cards from both groups together and shuffle. • Deal cards into two piles to mimic random assignment. • Record the difference in means for the two groups.	Applet: Randomization test for two means

11/9	z-test for difference of proportions	Distracted drivers (Starnes et al., 2013)	48 subjects were randomly assigned to two groups. One group drove in a simulator while talking on a cell phone, and the other group drove in a simulator while talking to a passenger. One outcome of interest was whether the driver would remember to stop at a rest area that was specified by researchers before the simulation started. The number who remembered in each group was recorded. Do the results of the study provide convincing evidence that talking on a cell phone is more distracting than talking to a passenger?	Chance device: cards (two colors) • Let one color represent drivers who stopped and one color represent drivers who didn’t stop. • Put cards from both groups together and shuffle. • Deal cards into two piles to mimic random assignment. • Record the difference in the proportion who stopped for the two groups.	Applet: Randomization test for categorical response

11/16	N/A	Hot hand (Starnes et al., 2013)	A basketball announcer believes that a certain player is streaky. That is, the announcer believes that if the player makes one shot then he is more likely to make his next shot. As evidence he points to a recent game where the player took 30 shots and had a streak of 7 shots made in a row. Assume the player makes 2/3 of his shots. Is this convincing evidence of "streakiness"?	Chance device: dice • Let 1-4 represents shots that he made. Let 5-6 represent shots he missed. • Roll die 30 times to represent a game with 30 shots. • Record the longest streak for each game.	None
11/18	N/A	Picking teams (Starnes et al., 2013)	At a department party, 18 students in the mathematics/statistics department at a university decide to play a trivia game. 12 of the students are math majors and 6 are stats majors. To divide into two teams of 9, one of the professors put all the players' names in a hat and drew out 9 players to form one team, with the remaining 9 players forming the other team. The players were surprised when one team was made up entirely of math majors. Does this outcome provide convincing evidence that the names weren't mixed well in the hat?	Chance device: cards • Write M on 12 cards to represent math majors and S on 6 cards to represent stat majors. • Shuffle the cards then draw 9 card to represent the selection of a team. • Record the number of math majors on the team that had the higher number of math majors.	TI-84 Plus: • Let 1-12 represent math majors and 13-18 stat majors • Use randIntNoRep to shuffle the integers • Students corresponding to first 9 integers are form a team.

11/18	Exact binomial test	1 in 6 wins (Starnes et al., 2013)	A soft drink company printed a message on the inside of each bottle cap. Some of the caps said, “Please try again!” while others said, “You’re a winner!” The company advertised the promotion with the slogan “1 in 6 wins a prize.” 7 friends each buy a bottle, and 3 of them won a prize. Is this convincing evidence that the company’s claim inaccurate?	Chance device: dice • Let 1-5 represents a loss and let 6 represent a win. • Roll die 7 times. • Record the number of wins.	Applet: One proportion inference
12/2	z-test for difference of proportions	Dolphin therapy (Zieffler & Catalysts for Change, 2013)	30 subjects with a clinical diagnosis of depression were randomly assigned to one of two treatment groups. Both groups engaged in swimming and snorkeling each day, but one group did so in the presence of bottlenose dolphins and the other did not. 10 of 15 subjects in the dolphin therapy group showed substantial improvement, compared to 3 of 15 subjects in the control group. Does these results provide convincing evidence that dolphin therapy is effective?	Chance device: cards (two colors) • Let one color represent subjects who improved and one color represent subjects who didn’t. • Put cards from both groups together and shuffle. • Deal cards into two piles to mimic random assignment. • Record the difference in the proportion who improved for the two groups.	Applet: Dolphin study applet

1/15 – 1/29: Simulations were used extensively to develop understanding of sampling distributions. However, these activities did not require the full logic of inference; specifically, students were not required to compare the results of a study to a simulated sampling distribution. Thus, they are not included in this appendix.					
2/12	z-test for one proportion	Facial prototyping (Tintle et al., 2013) Ch. 1 available here	Students are shown two faces and asked to identify which one was Tim and which one was Bob. The proportion of students who identify Tim as the man on the left is recorded. Do these data provide convincing evidence that the class is using facial prototyping (not just guessing)?	Chance device: coins • Heads – identifies the man on the left as Tim • Tails – does not identify the man on the left • Flip coin once for each student in the class. • Record the proportion who identify the man on the left as Tim	Applet: One proportion inference
2/22	z-test for one proportion	Innocent until Proven Guilty (Case & Whitaker, 2016)	In a two-player trivia game, a spinner is used to decide which player gets the first shot at answering the question. Player A had access to the spinner before the game, and player B suspects that he may have tampered with it to get more chances at answering questions. In groups, students test the spinner to see if they can convict player A of cheating.	Chance device: spinner Three spinners are used each with a different fraction corresponding to player A. See article for more details.	Tinkerplots: Model is available here

2/29	z-test for difference of proportions	Is yawning contagious? (Starnes et al., 2013) A short version of the episode is available here.	The *Mythbusters* team assigned 50 subjects to two groups. Two-thirds of the subjects were given a yawn seed; that is the experimenter yawned in the subject's presence. The remaining subjects were given no yawn seed. Of the 34 subjects who received a yawn seed, 10 yawned. Of the 16 who received no yawn seed, 4 yawned. Do these results provide convincing evidence that yawning is contagious?	Chance device: cards (two colors) • Let one color represent subjects who yawned and one color represent subjects who didn't. • Put cards from both groups together and shuffle. • Deal cards into two piles to mimic random assignment. • Record the difference in the proportion who yawned for the two groups.	Applet: Randomization test for categorical response
3/4	z-test for difference of means	Sleep deprivation (revisited)			

3/14	chi-square goodness of fit test	Favorite subject (original activity)	A Gallup survey asked American teens about their favorite subject in school; assume these percentages are population values. Compare these percentages to data we collected through a SRS of students at this school (n=42). Do our data provide convincing evidence that the distribution of favorite subject at this school is different from the Gallup results?	Student designed physical simulation but it was not carried out. One possible chance device: spinner • Areas corresponding to the Gallup results. • Spin 42 times. • Calculate χ^2 statistic for each trial.	Tinkerplots
4/4	t-test for the slope of the regression line	Seat location (Starnes et al., 2013)	An AP Statistics teacher randomly assigned 30 students to seat locations in his classroom (rows 1-7) for a particular chapter and recorded the test score for each student at the end of the chapter. The slope of the regression line relating score to row was -1.12. Does this provide convincing evidence that sitting closer causes higher achievement?	Chance device: cards • Write scores on cards. • Deal into seven rows to mimic random assignment. • Calculate the slope of the regression line relating score to row.	Applet: Analyzing two quantitative variables

5/6	Exact binomial test	Juice preferences AP Exam 2006B #6	Sunshine Farms wants to know whether there is a difference in consumer preference for two new juice products – Citrus Fresh and Tropical Taste. In an initial blind taste test, 8 randomly selected consumers were given unmarked samples of two juices; of these, 6 preferred Tropical Taste. Does this provide convincing evidence of consumer preference?	Chance device: coins • Heads – preference for Tropical Taste • Tails – preference for Citrus Fresh • Flip coin once for each subject in the original taste-test. • Record the number who preferred Tropical Taste.	Applet: One proportion inference

LIST OF REFERENCES

American Statistical Association. (2005). Guidelines for assessment and instruction in statistics education: College report. American Statistical Association. Retrieved from http://www.amstat.org/education/gaise/GaiseCollege_full.pdf

American Statistical Association. (2016). Guidelines for assessment and instruction in statistics education (GAISE) college report. Retrieved from http://www.amstat.org/education/gaise/collegeupdate/GAISE2016_DRAFT.pdf

Aquilonious, B. C., & Brenner, M. E. (2015). Students' reasoning about p-values. *Statistics Education Research Journal*, *14*(2), 7–27.

Aspinwall, L., & Tarr, J. E. (2001). Middle school students' understanding of the role sample size plays in experimental probability. *The Journal of Mathematical Behavior*, *20*(2), 229–245.

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). Kluwer Academic Publishers. Retrieved from http://link.springer.com/content/pdf/10.1007/1-4020-2278-6_1.pdf

Case, C., & Whitaker, D. (2016). Innocent until Proven Guilty. *Mathematics Teacher*, *109*(9), 686–692.

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

Chance, B., & McGaughey, K. (2014). Impact of a simulation/randomization-based curriculum on student understanding of p-values and confidence intervals. In *Proceedings of the 9th International Conference on Teaching Statistics* (Vol. 9). Retrieved from http://icots.info/9/proceedings/pdfs/ICOTS9_6B1_CHANCE.pdf

Chance, B., & Rossman, A. (2006). Using simulation to teach and learn statistics. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics*. Salvador, Brazil: ISI & IASE. Retrieved from http://www.ime.usp.br/~abe/ICOTS7/Proceedings/PDFs/InvitedPapers/7E1_CHAN.pdf

Charmaz, K. (2014). *Constructing grounded theory* (2nd ed.). London, England: SAGE.

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

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

College Board. (2010). AP Statistics Course Description. Retrieved from http://media.collegeboard.com/digitalServices/pdf/ap/ap-statistics-course-description.pdf

Conrad, C. (1982). Grounded theory: An alternative approach to research in higher education. *Review of Higher Education*, *5*(4), 239–249.

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, *13*(1), 3–21.

Creswell, J. W. (2013). *Qualitative inquiry and research design: Choosing among five approaches*. Thousand Oaks, CA: SAGE.

delMas, R. (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). New York, NY: Kluwer Academic Publishers. Retrieved from http://link.springer.com/content/pdf/10.1007/1-4020-2278-6_4.pdf

delMas, R., Garfield, J., & Chance, B. (1999). Assessing the effects of a computer microworld on statistical reasoning. *Journal of Statistics Education*, *7*(3).

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

Dey, I. (1999). *Grounding grounded theory*. San Diego, CA: Academic Press.

Franklin, C. A., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). Guidelines for assessment and instruction in statistics education (GAISE) report. American Statistical Association. Retrieved from www.amstat.org/education/gaise

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

Gordon, S., Reid, A., & Petocz, P. (2010). Qualitative approaches in statistics education research. *Statistical Education Research Journal*, *9*(2), 2–6.

Gould, R., Davis, G., Patel, R., & Esfandiari, M. (2010). Enhancing conceptual understanding with data driven labs. In *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics/*. Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://icots.info/icots/8/cd/pdfs/contributed/ICOTS8_C208_GOULD.pdf

Groth, R. E. (2010). Situating qualitative modes of inquiry within the discipline of statistics education research. *Statistics Education Research Journal*, *9*(2), 7–21.

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

Hodgson, T., & Burke, M. (2000). On simulation and the teaching of statistics. *Teaching Statistics*, *22*(3), 91–96.

Holcomb, J., Chance, B., Rossman, A., & Cobb, G. (2010). Assessing student learning about statistical inference. In C. Reading (Ed.), *Proceedings of the 8th International Conference on Teaching Statistics*. Ljubjana, Slovenia: ISI & IASE. Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/icots8/ICOTS8_5F1_CHANCE.pdf

Holcomb, J., Chance, B., Rossman, A., Tietjen, E., & Cobb, G. (2010). Introducing concepts of statistical inference via randomization tests. In C. Reading (Ed.), *Proceedings of the 8th International Conference on Teaching Statistics*. Ljubjana, Slovenia: ISI & IASE. Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/icots8/ICOTS8_8D1_HOLCOMB.pdf

Johnson, T., & Lesh, R. A. (2003). A models and modeling perspective on technology-based representational media. In R. Lesh & H. M. Doerr (Eds.), *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Kalinowski, P., Lai, J., Fidler, F., & Cumming, G. (2010). Qualitative research: An essential part of statistical cognition research. *Statistics Education Research Journal*, *9*(2), 22–34.

Konold, C. (1994). Understanding probability and statistics through resampling. In L. Brunelli & G. Cicchitelli (Eds.), *Proceedings of the First Scientific meeting of the International Association for Statistical Education* (pp. 255–263). Perugia, Italy: University of Perugia.

Konold, C., & Miller, C. (2011). Tinkerplots (Version 2). Emeryville, CA: Key Curriculum Press.

Kvale, S. (1996). *InterViews: An introduction to qualitative research interviewing*. Thousand Oaks, CA: SAGE.

Lane-Getaz, S. J. (2007). *Development and Validation of a Research-based Assessment: Reasoning about P-values and Statistical Significance*. University of Minnesota. Retrieved from https://www.stat.auckland.ac.nz/~iase/publications/dissertations/07.Lane-Getaz.Dissertation.pdf

Lehrer, R., & Schauble, L. (2006). Cultivating model-based reasoning in science education. In R. K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences* (pp. 371–405). New York, NY: Cambridge University Press.

Lehrer, R., & Schauble, L. (2007). Contrasting emerging conceptions of distribution in contexts of error and natural variation. *Thinking with Data*, 149–176.

Lesh, R. A., & Doerr, H. M. (2000). Symbolizing, communicating, and mathematizing: Key componenets of models and modeling. In P. Cobb, E. Yackel, & K. McClain (Eds.), *Symbolizing and communicating in mathematics classrooms: Perspectives on discourse, tools, and instructional design*. Mahwah, N.J.: Routledge.

Lesh, R. A., & Doerr, H. M. (2003a). *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching.* Mahwah, NJ: Lawrence Erlbaum Associates.

Lesh, R. A., & Doerr, H. M. (2003b). Foundations of a models and modeling perspective on mathematics teaching, learning, and problem solving. In R. Lesh & H. M. Doerr (Eds.), *Beyond constructivism: Models and modeling perpectives on mathematics problem-solving, learning, and teaching* (pp. 3–34). Mahwah, NJ: Erlbaum.

Lewis, C., Perry, R., & Murata, A. (2006). How should research contribute to instructional improvement? The case of lesson study. *Educational Researcher*, *35*(3), 3–14.

Lock, R., Lock, P. F., Morgan, K. L., Lock, E., & Lock, D. (2014). Intuitive introduction to the important ideas of inference. In K. Makar (Ed.), *Proceedings of the Ninth International Conference on Teaching Statistics*. ISI & IASE. Retrieved from http://icots.info/icots/9/proceedings/pdfs/ICOTS9_4A3_LOCK.pdf

Maher, C. A., & Sigley, R. (2014). *Encyclopedia of Mathematics Education*. (S. Lerman, Ed.). Dordrecht: Springer Netherlands.

Malone, C., Gabrosek, J., Curtiss, P., & Race, M. (2010). Resequencing topics in an introductory applied statistics course. *The American Statistician*, *64*(1), 52–58.

Mills, J. D. (2002). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education*, *10*(1), 1–20.

Mittag, K. C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher*, *29*(4), 14–20.

Moore, D. S. (1990). Uncertainty. In L. Steen (Ed.), *On the shoulders of giants: new approaches to numeracy*. Washington, D.C.: National Academy Press.

National Governors Association Center for Best Practice, & Council of Chief State School Officers. (2010). *Common Core State Standards*. Washington, D.C.: Author.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy.

Pfannkuch, M. (2005). Probability and statistical inference: How can teachers enable learners to make the connection? In G. A. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 267–293). New York, NY: Springer.

Reaburn, R. (2014). Introductory statistics course tertiary students' understanding of p-values. *Statistics Education Research Journal*, *13*(1), 53–65.

Rossman, A. J., & Chance, B. L. (2014). Using simulation-based inference for learning introductory statistics. *Wiley Interdisciplinary Reviews: Computational Statistics*, *6*(4), 211–221.

Saldanha, L., & Thompson, P. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics*, *51*(3), 257–270.

Schoenfeld, A. H. (1985). Making sense of "out loud" problem-solving protocols. *Journal of Mathematical Behavior*, *4*, 171–191.

Schoenfeld, A. H. (2007). Method. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 69–107). Charlotte, NC: Information Age Publishing.

Seel, N. M. (2014). Model-Based Learning and Performance. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of Research on Educational Communications and Technology* (pp. 465–484). New York, NY: Springer New York.

Starnes, D. S., Tabor, J., Yates, D., & Moore, D. S. (2013). *The Practice of Statistics* (5th ed.). W.H. Freeman.

Stohl, H., & Tarr, J. E. (2002). Developing notions of inference using probability simulation tools. *The Journal of Mathematical Behavior*, *21*(3), 319–337.

Taber, K. S. (2000). Case studies and generalizability: grounded theory and research in science education. *International Journal of Science Education*, *22*(5), 469–487.

Taylor, L., & Doehler, K. (2015). Reinforcing Sampling Distributions through a Randomization-Based Activity for Introducing ANOVA. *Journal of Statistics Education*, *23*(3). Retrieved from http://www.amstat.org/publications/jse/v23n3/taylor.pdf

Tintle, N., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (2013). *Introduction to Statistical Investigations*. Hoboken, NJ: John Wiley and Sons.

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

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), n1.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223–248.

Zieffler, A., & Catalysts for Change. (2013). *Statistical thinking: A simulation approach to uncertainty* (2nd ed.). Minneapolis, MN: Catalyst Press.

Zieffler, A., delMas, R., Garfield, J., & Brown, E. (2014). The symbiotic, mutualistic relationship between modeling and simulation in developing students' statistical reasoning about inference and uncertainty. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Proceedings of the 9th International Conference on Teaching Statistics, ICOTS*. Voorburg, The Netherlands: ISI & IASE. Retrieved from http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_8B1_ZIEFFLER.pdf

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

BIOGRAPHICAL SKETCH

Catherine graduated from the University of Florida in the fall of 2016 with a Ph.D in curriculum and instruction. Her major area of concentration was statistics education and her minor was research and evaluation methodology. As part of her research fellowship, she worked with Dr. Tim Jacobbe on the LOCUS project – an NSF-funded grant to develop an assessment of conceptual understanding in statistics.

While earning her master's in statistics at UF, Catherine taught several introductory statistics course for the Department of Statistics. While earning her doctorate, she taught a mathematics course for pre-service elementary teachers and an AP Statistics course at P.K. Yonge Developmental Research School. For her effects working with high school students at P.K. Yonge alongside Douglas Whitaker and Steven Foti, she received an I-Cubed Graduate Student Mentoring Award in 2014. Catherine was also an author of the *Statistical Education of Teachers* report – a recommendations document commissioned by the American Statistical Association.

In July 2016, Catherine and her husband, Adam, moved to Athens, Georgia. Catherine accepted a position at the University of Georgia.