

A FRAMEWORK TO CHARACTERIZE STUDENT DIFFICULTIES IN LEARNING INFERENCE FROM A SIMULATION-BASED APPROACH

CATHERINE CASE
University of Georgia
catherine.case@uga.edu

TIM JACOBBE
University of Florida
jacobbe@coe.ufl.edu

ABSTRACT

Although hypothesis testing is ubiquitous in data analysis, research suggests it is commonly misunderstood. Simulation-based inference methods have potential to make student thinking visible, thus providing a valuable lens to analyze developing conceptions about inference. This paper identifies difficulties made visible through simulation-based methods and introduces a framework to characterize the conceptions behind those difficulties. Using the framework, difficulties 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 a 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 were collected, and the hypothetical perspective where the null hypothesis is assumed to be true.

Keywords: *Statistics education research; Simulation; Inferential reasoning; Introductory statistics*

1. INTRODUCTION

In its list of goals for students in introductory statistics courses, the *Guidelines for Assessment and Instruction in Statistics Education (GAISE): College Report* (GAISE College Report American Statistical Association Revision Committee, 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). If people have not experienced sampling variability in their education, they instinctively look for deterministic causes rather than consider chance variation (Wild & Pfannkuch, 1999), which may lead to over-interpretation of results. Further, research suggests that many students who complete an introductory statistics class still hold misconceptions about inference (delMas, Garfield, Ooms, & Chance, 2007). Even among university faculty and research professionals, many lack a full understanding of statistical significance and p -values (Haller & Krauss, 2002; Mittag & Thompson, 2000; Nickerson, 2000). Widespread misuse and misunderstanding of inference “affects not only research, but research funding, journal practices, career advancement, scientific education, public policy, journalism, and law” (Wasserstein & Lazar, 2016, p. 8). Thus, promoting rich understanding of inference must be a high priority for teachers of statistics.

Many statistics educators (e.g., Chance & Rossman, 2006; Cobb, 2007; Lock, Lock, Morgan Lock, Lock, & Lock, 2014; Pfannkuch, 2005) believe that simulations have the potential to develop 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 theory-based inference (GAISE College Report ASA Revision Committee, 2016; Rossman & Chance,

2014). Adoption of these methods has been 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. Developers of curricula that employ simulations as the primary means of teaching inference have published evaluations to suggest students in simulation-based courses compare favorably to students in theory-based courses (e.g., Garfield, delMas, & Zieffler, 2012; Tintle, Topliff, Vanderstoep, Holmes, & Swanson, 2012; Tintle, Vanderstoep, Holmes, Quisenberry, & Swanson, 2011).

Of course, a simulation-based approach does not eliminate all difficulties associated with learning inference. The logic of inference, whether it is introduced via simulations or via statistical theory, always invokes prerequisite concepts of variability, distributions, and sampling (Chance, delMas, & Garfield, 2004) and requires “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). When students carry out a well-rehearsed procedure—e.g., using a formula to calculate a test statistic, using a table to find a p -value, and comparing the p -value to a significance level to make a decision—problematic conceptions may go undiagnosed (Aquilonius & Brenner, 2015; Taylor & Doehler, 2015). 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.

The purpose of this study is to identify difficulties made visible through the use of simulation-based inference methods and to characterize the statistical conceptions behind those difficulties. The characterization of student conceptions necessitates a new two-dimensional framework for conceptualizing the logic of inference. The framework was constructed based on a review of the literature and a qualitative study of students who experienced both theory-based and simulation-based inference methods in instruction; examples from both the literature review and the current study will be included as illustrations.

Although the framework can be applied to inferential reasoning more generally, this article focuses on the difficulties that arise in learning statistical inference. The process of sound inferential reasoning using a simulation-based approach has been outlined elsewhere (e.g., Cobb, 2007; Garfield et al., 2012), but less has been written about the difficulties that must be overcome to reach that ideal. Awareness of common difficulties provides a valuable opportunity for teachers to recognize and challenge students’ emerging conceptions of inference.

2. SIMULATION-BASED INFERENCE

To provide background for the study, this section introduces simulation-based inference and presents the proposed advantages and empirical evaluations that have contributed to its recent popularity in introductory statistics classes. The section concludes with a list of difficulties related to simulation-based inference that have been identified in the literature; each will be explored more deeply later in the paper.

2.1. THEORY-BASED AND SIMULATION-BASED INFERENCE MODELS

Viewed through a *models and modeling* lens (Lesh & Doerr, 2003), theory-based 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. Defined broadly, statistical inference includes four main ideas: significance, estimation, generalizability, and causation (Rossman & Chance, 2014). The research presented here focuses on significance, so the term *inference* will be used to refer to hypothesis tests, including hypotheses about populations and hypotheses about causal relationships. This section presents the commonalities of inferential reasoning across simulation-based and theory-based approaches, borrowing language from statistics educators who have described hypothesis testing as a unified modeling process (e.g., Cobb, 2007; Garfield et al., 2012; Tintle et al., 2013).

The foundational problem of inference is variability. Because statistics vary from sample to sample, the observed outcome is not expected to reflect the population or true relationship perfectly. To account for imperfect correspondence between the sample statistic and the true parameter, students must

consider the distribution of statistics that could occur for a given parameter value. In hypothesis testing, a model is specified to approximate the variability in outcomes that would occur by chance if the null hypothesis were true. An observed statistic that would be unlikely to occur by chance provides evidence against the null hypothesis, and a p -value quantifies the likelihood that the observed statistic or one more extreme would occur if the null hypothesis were true. Thus, when the p -value is small, we reject the null hypothesis, ruling out a “just by chance” explanation for the observed outcome and concluding in favor of the alternative hypothesis (Cobb, 2007; Garfield et al., 2012; Tintle et al., 2013).

In the description above, the term *model* is defined broadly (Lesh & Doerr, 2003). Traditionally, statisticians used theoretical probability distributions to model the variability in sample statistics that would be expected to occur by chance if the null hypothesis were true. In this paper, hypothesis tests based on theoretical distributions (e.g., Normal distribution, t distribution, X^2 distribution) will be called *theory-based 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*.) Note that the term *simulate* is used differently here than in other disciplines like science. In other disciplines, simulation models are often constructed as a best-guess representation of reality and used to predict what would happen in the real-world. However, in hypothesis testing, simulations begin with the assumption that the null hypothesis is true. A detailed example of simulation-based inference is given in Section 4.3.

Though both theory-based and simulation-based methods embody the same logic of inference, they differ in the tools and representations employed. Lehrer and Schauble (2007) argue that “representational change both reflects and instigates new ways of thinking about the data” (p. 157); thus, it is important for statistics educators to understand how students interact with the various models and representations used in statistics instruction.

2.2. STUDENTS’ UNDERSTANDING OF STATISTICAL INFERENCE

Understanding of inference is central to statistical practice, but empirical studies suggest that inferential concepts present considerable difficulties to students. 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 (Aquilonius & Brenner, 2015; Holcomb, Chance, Rossman, & Cobb, 2010; Taylor & Doehler, 2015). A large-scale administration of the CAOS assessment to introductory statistics students from diverse institutions identified numerous inferential concepts that are commonly misunderstood (delMas et al., 2007). After a semester of instruction, fewer than 60% of students correctly answered items measuring the following outcomes: understanding that statistics from small samples vary more than statistics from large samples (only 31.9% answered correctly on the posttest); ability to recognize an incorrect interpretation of a p -value (probability that a treatment is effective) (52.7%); understanding of the logic of a significance test when the null hypothesis is rejected (52.0%); ability to recognize a correct interpretation of a p -value (54.5%); and ability to recognize an incorrect interpretation of a p -value (probability that a treatment is not effective) (58.6%). Based on a literature review of ten empirical studies, Lane-Getaz (2007) categorized difficulties related to understanding statistical significance into four categories: misunderstanding terminology and basic concepts, confusing relationships between inferential concepts, misapplying the logic of statistical inference, and misinterpreting the p -value as the probability of the truth or falsity of hypotheses.

Chance, delMas, and Garfield (2004, p. 295) attribute poor understanding of inference to “the notoriously difficult, abstract topic of sampling distributions.” 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. (p. 7)

Further, in traditional approaches to teaching 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).

2.3. 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). Because 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 theory-based inference cannot be introduced without the machinery of theoretical sampling distributions (Holcomb, Chance, Rossman, Tietjen, & Cobb, 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 generalizes easily to a large number of settings; in theory-based inference, “there are so many variations that it is hard for students to recognize and appreciate the unifying themes” (Cobb, 2007, p. 6). Third, a simulation-based approach 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 also modernizes the content of the introductory statistics courses to reflect technological advances (Cobb, 2007; Holcomb, Chance, Rossman, Tietjen, et al., 2010).

2.4. EMPIRICAL COMPARISONS OF CURRICULA

In addition to philosophical arguments, researchers have begun to empirically evaluate the impact of simulations on students’ understanding of inference. Evaluations of simulation-based curricula find student performance is similar for students who study simulation-based and theory-based 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).

The studies referenced above feature quantitative analysis of student performance on multiple-choice assessments, specifically CAOS (delMas et al., 2007)—a validated assessment designed to measure students’ conceptual understanding after a first course in statistics. Because the CAOS assessment is not specific to simulation-based inference, it provides an appropriate metric for comparing the two approaches. However, it cannot provide specific information about difficulties made visible through simulation-based inference.

2.5. COMMON STUDENT DIFFICULTIES IDENTIFIED IN THE LITERATURE

Large-scale quantitative evaluations have not provided theory to explain how novices employ simulation-based inference, but developers and users of these curricula have shared brief recommendations regarding difficulties that arise, often disseminated in the form of conference papers and presentations. Common difficulties and conceptions that have been reported include the following:

- 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)

In addition to the challenges mentioned above, Chance and McGaughey (2014) warn about “the difficulty students may have in the transition from one 50/50 proportion to other scenarios, including the distinction between sampling and assignment” (p. 6). As part of a larger study, the authors have explored various ways that student-designed models can reveal student thinking in more complex data scenarios, but this paper restricts attention to difficulties that are relevant to all simulation-based tests, including tests of a single proportion.

2.6. RESEARCH QUESTION

The literature reviewed in this section provides background for study of student difficulties and conceptions, but the existing empirical studies, which feature quantitative analysis of student performance on summative assessments, have not provided theory to explain how novices employ theory-based and simulation-based inference models. Further, although simulation-based inference is included in the Common Core State Standards for Mathematics (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) and many high school students study inference in Advanced Placement (AP) Statistics courses, a review of the literature revealed little about high school students’ understanding of inference. This paper forms part of a larger qualitative study that aims to address those gaps in the literature by pursuing the following research question: How do students use traditional inference models and simulation-based inference models to understand inference? This broad research interest was refined in two sub-questions:

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

This paper focuses on the first sub-question. Likewise, the description of methods will emphasize data collection and analysis processes relevant to that question.

3. METHODS

3.1. CONTEXT

This study was situated in an AP Statistics course taught by the first author at a public school in the southeastern United States. The AP Program offers opportunities for high school students (generally ages 14–18) to take college-level classes, and most four-year colleges and universities in the United States offer students course credit and/or placement based on AP Exam scores (College Board, 2010). The AP Statistics curriculum includes four major topics (College Board, 2010): data analysis and exploration (20-30% of the exam), study design (10-15% of the exam), probability and simulation (20-30% of the exam), and statistical inference (30-40% of the exam). In addition to the prescribed AP Statistics curriculum, which relies on theory-based inference methods, the course under study regularly incorporated simulation-based methods in an effort to improve conceptual understanding of inference. This complementary approach was supported by the course textbook, *The Practice of Statistics* (Starnes, Yates, & Moore, 2012).

Theory-based inference requires substantial prerequisite knowledge, including knowledge of probability and theoretical sampling distributions. For this reason, introductory statistics courses often consist of three major units: (1) descriptive statistics and study design; (2) probability and sampling distributions; and (3) statistical inference (Malone, Gabrosek, Curtiss, & Race, 2010). However, simulation-based inference can be introduced much earlier, because it requires less prerequisite knowledge (Cobb, 2007). In the course under study, simulation-based inference activities were incorporated throughout the year beginning on the first day of class, with theory-based inference introduced in the final third of the course. 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 their own simulations using physical chance devices and applets. However, because the AP

Statistics course description emphasizes theory-based inference, students still had considerably more experience with theory-based methods by the end of the school year.

To help students recognize the unified modeling process behind statistical inference, the teacher adopted the 3S Strategy presented in the *Introduction to Statistical Investigations* curriculum (Tintle et al., 2013):

- **Statistic:** Compute the statistic from the observed sample data.
- **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.
- **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... (p. 28)

Theory-based inference was first introduced as a modification to the 3S Strategy, where the simulation step was replaced by use of a theoretical sampling distribution. For each new data type (e.g., one proportion, two means, etc.), students first applied the 3S strategy, using simulations to test for statistical significance. Then they transitioned to theory-based significance tests, calculating the test statistic using a formula and finding the p -value using a cumulative density function in the calculator. After using these tools a few times, students were introduced to functions on the TI-84 Plus calculator that calculate test statistics and p -values using summary statistics or raw data as inputs. Throughout the year, connections between simulation-based inference and theory-based inference were made explicit, and assessments prompted students to reflect on these connections.

3.2. DATA COLLECTION

The data for this study were collected from AP Statistics students taught by the first author in two different years. In the pilot study, seven students—selected to represent a range of class achievement, as measured by cumulative grades near the end of the year—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 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 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 Individual interview tasks prompted students to conduct hypothesis tests to draw conclusions about the results of research studies. These tasks are reproduced in Figure 1; Task 1 was proposed by Holcomb, Chance, Rossman, Tietjen, et al. (2010), and Task 2 was included in *The Practice of Statistics, Teacher’s Edition* (Tabor, Starnes, Yates, & Moore, 2012). Each student was given one of the two tasks and was asked to apply two different methods—a theory-based test and a simulation-based test—in the 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, fourteen individual interviews were conducted—seven in each year of the study.

Group interviews Additionally, all eleven students enrolled in the class during the second year of the study were invited to participate in group interviews. Ten of these students were interviewed in pairs. (One was unable to participate.) Similar in structure to the individual interviews, students were given the results of a study and were asked to work together to decide whether the study provided convincing evidence against the null hypothesis. The task, original to this study, is shown in Figure 1 (Task 3).

Task 1: Helper vs. Hinderer (Holcomb, Chance, Rossman, Tietjen, et al., 2010, pp. 1–2)

In a study reported in *Nature* (Hamlin, Wynn, and Bloom, 2007), 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 (Tabor et al., 2012, p. 245)

In a study reported in the *New England Journal of Medicine* (Knapp & Fitzgerald, 1989), researchers investigated whether fish oil can help reduce blood pressure. Fourteen 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 1. Interview tasks

Schoenfeld (1985) suggested 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 make the reasoning behind their decisions more visible (Schoenfeld, 1985).

Student work Formative assessments and exam items were used 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.

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 informed the research questions.

3.3. 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—for example, *proposing a chi-square test, reading calculator output, deciding to reject the null based on a rule, using equal numbers of cards to represent yes/no*. Each discrete action on the part of the student constituted a segment of data, and in keeping with the recommendations put forth by Charmaz, gerunds were used as initial codes to focus on students' actions and stay close to the data, encouraging the researcher to “begin analysis from [the participants'] perspective” (Charmaz, 2014, p. 121).

After comparing these initial codes to the data and looking for patterns in the codes across interviews, focused codes were constructed inductively. This paper focuses on codes used to label common difficulties—for example, *conclusions based on one simulated sample, conflation of number of trials and sample size, and misuse of the sampling distribution to estimate the p-value*. Focused codes facilitated comparisons across tasks, across students, and across inferential approaches. In particular, data coded for common difficulties were subjected to two types of systematic comparisons. First, data assigned the same focused code were compared across participants. Second, data coded for common difficulties were compared to other work produced by the same participant.

As an illustration of the data analysis process, consider the following example. On an exam, students were asked to complete the task below, which was modified from a textbook exercise (Starnes et al., 2012). The exam item is followed in the text by Isabella's response.

Biologists conducted a study in an enclosed outdoor space with a piece of shore whose area was made up of 56% sand, 29% mud, and 15% rocks. The biologists chose 200 seagulls at random. Each seagull was released into the outdoor space on its own and observed until it landed somewhere on the piece of shore. In all, 128 seagulls landed on sand, 61 landed in mud, and 11 landed on rocks. 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, ...)

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 χ^2 -GOF.

Isabella's work can be broken into five segments, each assigned an initial code: *using a spinner to model the breakdown of the outdoor space, choosing a number of spins to match the sample size, organizing the simulated data, calculating expected counts, and using a calculator to conduct a goodness-of-fit test on simulated data*. In the next round of coding, her full response was assigned a focused code: *combination of simulation-based and theory-based approaches*. It was then compared to other data with the same focused code, including student work where a group of students used an applet to simulate a sampling distribution then used a *z*-statistic and calculator to find a *p*-value on a formative assessment and a passage from the teacher-researcher's journal (reproduced below).

Journal: ...Ultimately, Eva was able to help Alicia and Grace understand how collecting the slopes would lead to a simulated distribution. At the strength of evidence step, Ryan mentioned that we needed “something-something cdf” in the calculator to get a *p*-value which could be compared to an alpha level. Until I gave leading prompts, no one mentioned that this represented a mix of approaches...

Finally, Isabella’s response to the exam item was compared to other work she produced on exams, in class, and in interviews to better understand how her approach to this problem fit with her broader conceptions of inference. As illustrated above, these comparisons aimed to contextualize errors to better understand students’ underlying conceptions. Ultimately, the comparisons revealed patterns of student difficulties, which provided the basis for a new conceptualization of the logic of inference. The framework is presented in the next section.

4. CONCEPTUALIZING THE LOGIC OF INFERENCE

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

4.1. TWO PERSPECTIVES: REAL WORLD AND HYPOTHETICAL

Note that coordination of the components mentioned by Zieffler et al. (2008) requires shifting between two perspectives. The sample data were produced by a randomized process—for example, random sampling or random assignment—in the real world. On the other hand, the distribution of the statistic, or the sampling distribution, was 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.

4.2. THREE LEVELS: POPULATION/TRUE RELATIONSHIP, SAMPLE, AND SAMPLING DISTRIBUTION

The goal of statistical inference is to use a statistic calculated from a particular sample in the real world to draw conclusions about a larger population or an underlying causal relationship. Because statistics vary and the observed data are not expected to reflect the population parameter or true relationship perfectly, statistical inference is more complex than *modus tollens* logic in a deterministic scenario. To account for sample-to-sample variability, students must consider the distribution of statistics that could be generated if the null hypothesis were true. 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 (the sampling distribution). If the real-world statistic (from the sample) is extreme in comparison to the distribution of statistics that would occur under the null hypothesis (sampling distribution), then the null hypothesis is rejected, and conclusions can be drawn about the population or true relationship.

4.3. ILLUSTRATION OF THE FRAMEWORK

Figure 2 illustrates the two-dimensional framework in the context of Task 2, which tested the effect of fish oil on blood pressure. Note that the cells of Figure 2 are labeled, and these labels are used throughout the paper for clarity. Rossman/Chance applets were used to create the graphical displays: <http://www.rossmanchance.com/applets/>

On average, men in the fish oil group saw larger reductions in blood pressure than men in the regular oil group (Figure 2, cell R2); the observed difference of means was $\bar{x}_{fish} - \bar{x}_{regular} = 7.714$.

However, these sample data may not perfectly reflect the unknown, real-world relationship between fish oil and blood pressure (R1). It is still possible that the type of oil has no effect on blood pressure, and the population mean blood pressure reductions for those who consume fish oil and those who consume regular oil are equal; this possibility is the null hypothesis (H1).

In order to use real-world empirical results to evaluate the null hypothesis, students must consider possible empirical results that could be generated if the null hypothesis were true, thus 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.

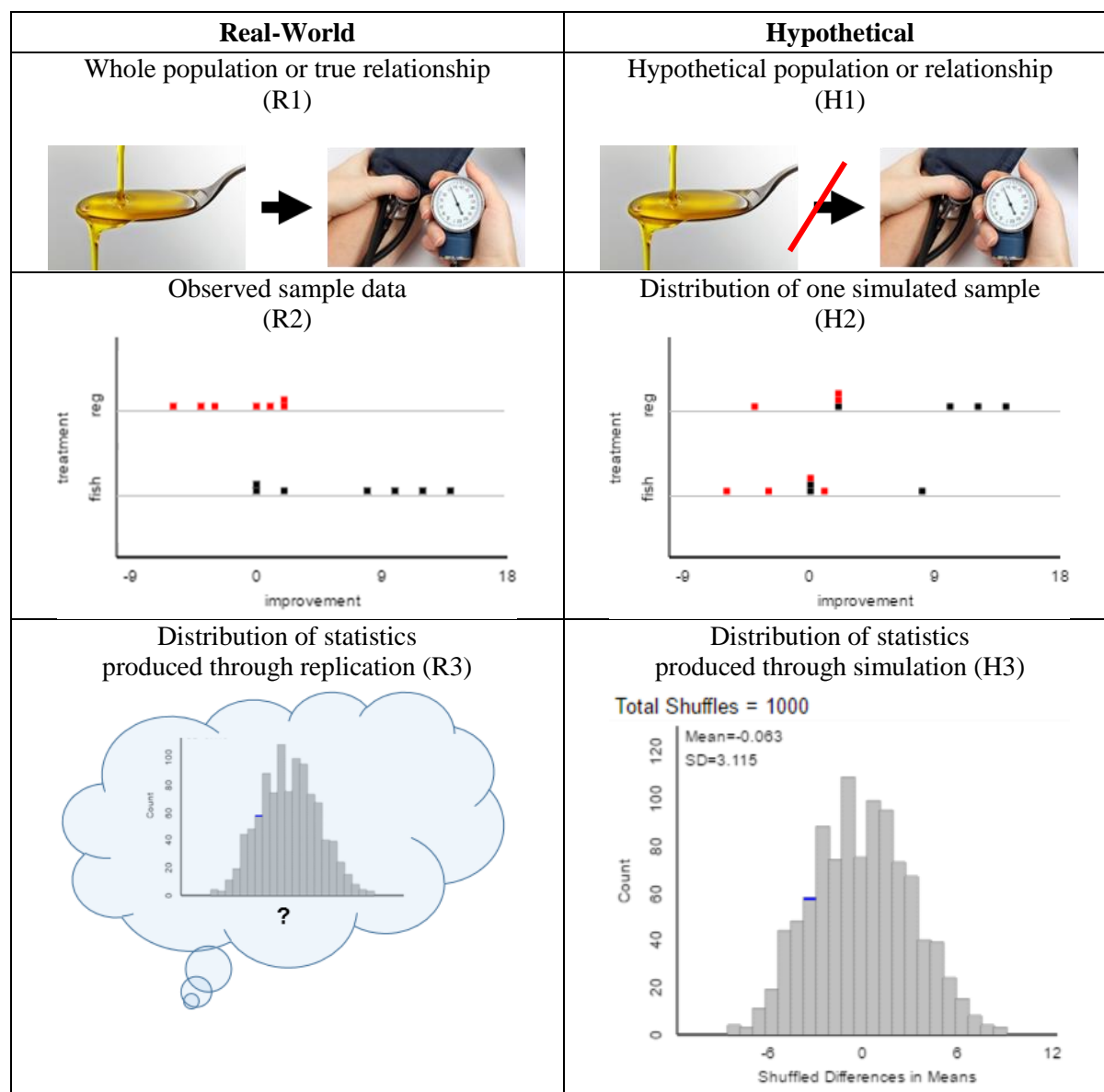


Figure 2. Levels and dimensions of inferential reasoning

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 and dealt into two piles, matching the original group sizes. The hypothetical dotplot (H2), shows the empirical results of one such simulated trial. Although this simulation model makes use of real-world data, it requires a hypothetical perspective, as it assumes that fish oil and regular oil have the same effect on blood pressure, and any differences in the sample means are determined by chance factors unrelated to treatment.

The simulation process is then repeated many times to create a distribution of summary statistics that shows which outcomes are typical under the null hypothesis. The hypothetical histogram (H3) 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.

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 or greater would be very unlikely to occur by random assignment if the two types of oil had the same effect on blood pressure, there is sufficient evidence to reject the null hypothesis that there is no difference in means.

Of the six elements shown in Figure 2, only the distribution of sample data (R2) is observable in the real world. The true, real-world relationship between fish oil and blood pressure is never fully known, even when the null hypothesis is rejected. Additionally, we never see a distribution of statistics produced through real-world replication of the study, though the concept of such a distribution can be a source of confusion for students (as described in the next section). Thus, many important elements for inferential reasoning are accessible only through models and their representations—a nontrivial challenge for students.

5. STUDENT DIFFICULTIES AND UNDERLYING CONCEPTIONS

The purpose of this study is to identify difficulties that arise among students who use simulation-based inference methods and to characterize the statistical conceptions behind those difficulties. After reviewing the literature and completing focused coding of the interview transcripts, student work, and teacher reflections, it became clear that many of the difficulties associated with simulation-based inference can be attributed to the challenge of coordinating multiple perspectives and levels simultaneously. This section describes difficulties that arise when student conflate or fail to make appropriate connections among the components of the framework shown in Figure 2.

5.1. DISTINGUISHING SAMPLES AND SAMPLING DISTRIBUTIONS

The literature reports that students often struggle to distinguish the distribution of the sample (R2) from the simulated sampling distribution (H3). In the context of a teaching experiment, Saldanha and Thompson (2002) investigated students' reasoning about samples and sampling distributions in a high school statistics class. As part of the instructional unit, students viewed computer simulations of repeated random sampling from a population. However, 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, students' conflation of samples with sampling distributions was apparent in their confusion of the sample size with the number of simulated trials. For example, when asked why we should repeat the simulation many times, students' answers were often ambiguous; advantages like "less variability" or "more accuracy" could be interpreted as advantages of a large sample size rather than a large number of trials. In other cases, they communicated their conceptions more explicitly. In the incident below, Isabella and Devon compare the sample size to the number of trials as they reason about the slightly different p -values produced by theory-based and simulation-based methods on the same inference task.

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

- Devon: Oh, exactly.
 Interviewer: Wait, what?
 Isabella: We had—like our [observed] 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.

Devon and Isabella reasoned that the simulation-based method (which involved a large number of simulated trials) was more reliable than the theory-based method where the sample size was smaller; that is, they compared the sample size in the observed data (R2) to the number of simulated samples (H3). Because “sample size” and “number of samples” sound similar and both are often represented by the letter N (lowercase and uppercase, respectively), this error may appear to be a typo on an in-class assessment. However, in this study, further questions and analysis often revealed the same conception described by Saldanha and Thompson (2002) and Rossman and Chance (2014).

After encountering both theory-based and simulation-based inference in instruction, some students in this study also experienced confusion about whether the Normality condition was related to sample size or the number of simulated samples. 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 using an applet. 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 3, this pattern is only 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.

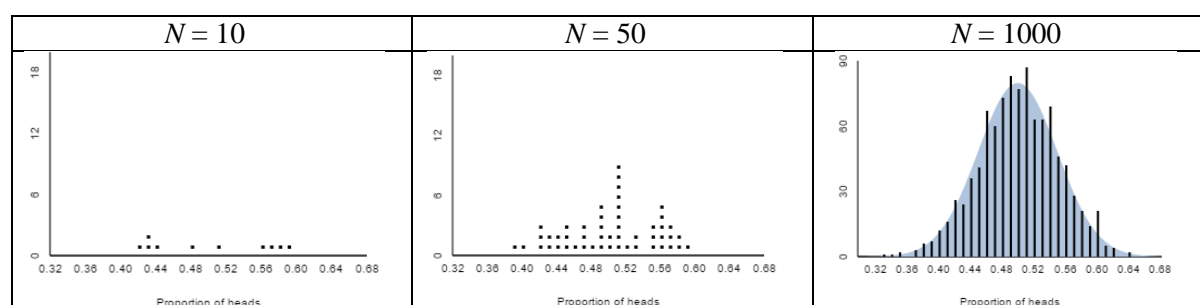


Figure 3. Simulated sampling distribution of a sample proportion

5.2. TRANSITIONING FROM THE SAMPLE LEVEL TO THE SAMPLING DISTRIBUTION LEVEL

The literature and the current study substantiate another difficulty related to coordinating levels. Some students can distinguish the units in the sample and the units in the sampling distribution while using a simulation applet, yet they struggle to envision the multi-level nature of inference when the applet is not open. Specifically, these students struggle to transition from the sample level (level 2) to the sampling distribution level (level 3). This issue was visible in student work in class and on exams, as well as in task-based interviews.

For example, in her individual interview, Eva designed a simulation using cards to re-randomize improvement scores in the blood pressure study (Task 2). When asked what she would do after dealing the cards into two groups, Eva pointed to the graph of the sample data (R2) 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, R2) ... you would put that on the graph. You would just graph what you got a bunch of times.
 Interviewer: So you would have a bunch of graphs that look like this?
 Eva: No, it would be the same graph.

- Interviewer: 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.
- Interviewer: 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 (H3) where the units on the graph are simulated statistics. However, later in the interview after opening the applet, Eva was able to reason with the graphical representation of simulated statistics, appropriately describing the units of the simulated distribution (H3) as a difference of sample means in context.

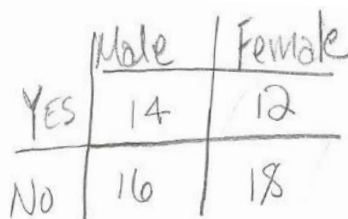
Zieffler, delMas, Garfield, and Brown (2014) report a similar phenomenon among students enrolled in the CATALST course—a college level course that uses simulations in *TinkerPlots* (Konold & Miller, 2005) to carry out statistical inference. In their study, a student designed a model and described simulation 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. In this interview, Eva was working with her partner Ryan to decide whether the data provided convincing evidence of response bias based on the gender of the interviewer (Task 3). They considered two approaches to compare the observed sample data (R2) with the results of a single simulated trial (H2). The data from the inference task are reproduced in Table 1.

Table 1. Data for group interview task (R2)

Feminist?	Gender of interviewer		Total
	Male	Female	
Yes	11	15	26
No	19	15	34
Total	30	30	60

In the incident that follows, Eva and Ryan had just simulated one trial using two colors of index cards to represent students who identified as feminist (answered “Yes” when asked by the interviewer) and students who did not (answered “No”). The results in Figure 4 show how many feminists and non-feminists were randomly assigned to the male and female interviewer in the first trial of their simulation.



	Male	Female
YES	14	12
NO	16	18

Figure 4. Student-generated table representing one simulated sample (H2)

- Ryan: Now we compare [the simulated data in Figure 4] to our results here [the observed data in Table 1]. 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 knew that their simulation represented a “by chance” explanation; that is, simulation models produce outcomes that would occur by chance if the null hypothesis were true. He treated the simulated counts as expected counts, because expected counts are also based on the null hypothesis. Then he pointed out that fewer people answered yes with a male interviewer than he would have expected based on the simulation. However, at this point in the interview, Ryan 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-squared test. Attempts to combine theory-based and simulation-based inference methods by calculating test statistics and p -values from simulated data were common among the students in this study, who were exposed to both theory-based and simulation-based tests in class. Students’ justifications for their choices also 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 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.
- Interviewer: So you’d do another t -test but on your simulated data?
- Libby: Yes.
- Interviewer: 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.

Although the data Libby used was produced by a physical simulation model of her own design, in that moment she failed to view her simulated data from the hypothetical perspective. Other issues related to coordinating perspectives are described in the next section.

5.3. DISTINGUISHING SIMULATION AND REPLICATION

Another set of difficulties stem from students imagining a distribution produced through simulation as a distribution produced through replication. In the real world, we never see a distribution of statistics produced by replicating a study many times, but it is common for students to conflate simulation with replication—unintentionally shifting from a hypothetical perspective to a real-world perspective. This conception manifests as a number of different issues that initially appear dissimilar.

One issue that can arise through simulations is the belief that multiple samples are always necessary. In an activity led by Hodgson and Burke (2000) to develop understanding of the Central Limit Theorem, students in an introductory statistics 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 required a hypothetical perspective where the parent population is somehow known, some students mistook the simulation process for “a real-world strategy for finding a population parameter” (Hodgson & Burke, 2000, p. 94).

Others have reported that this belief 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” (Rossman & Chance, 2014, p. 218). Confusion of simulation and replication occurred several times in the present study, even in interviews conducted at the end of the course. Students holding this conception treat simulated samples as a way to replicate the entire study. Thus, students may believe that flaws in the original study design can be corrected in simulated replications. For example, in her individual interview, Laura obtained substantially different p -values

from the theory-based and simulation-based approaches because of a calculation error. She offered the following explanation for the discrepancy:

Laura: In the [theory-based 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.

Conflation of simulation and replication can also lead to misapplications of the logic of inference. 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 took this approach as she considered whether 14 out of 16 babies choosing the helper toy provided convincing evidence of a genuine preference for the helper toy over the hinderer toy in Task 1 (testing the alternative hypothesis $p_{\text{helper}} > 0.5$.)

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.

Maria chose 8, which is the center of her null distribution of counts (Figure 5); that is, if the babies had no preference, 8 out of 16 would be expected to choose the helper. Though she had constructed the simulated distribution under the assumption that the null hypothesis was true (using a fair coin as a model), Maria interpreted the results as a lack of evidence against the null hypothesis.

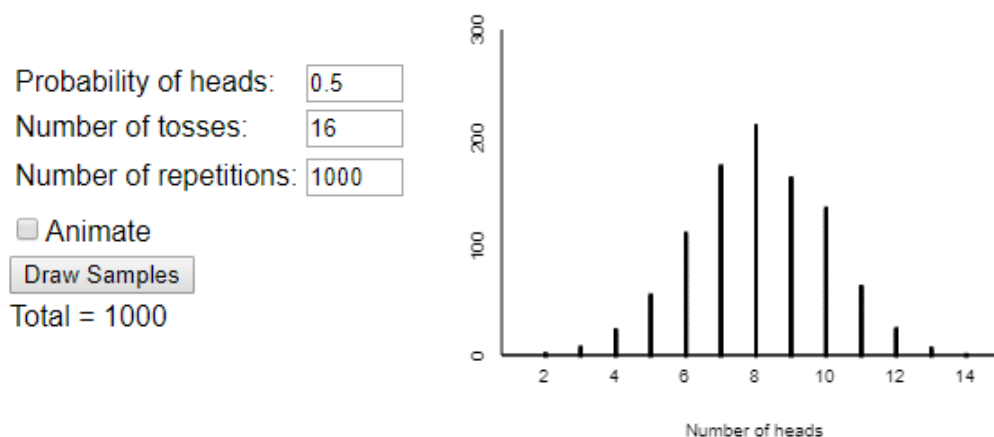


Figure 5. Simulated sampling distribution of the number who choose the helper

As described above, some students base their decision on the center of the sampling distribution, in which case failing to reject the null is a foregone conclusion. Other students who conflate simulation and replication do find the proportion of samples more extreme than the observed data to estimate a p -value. However, if that proportion is small, these students reason that the original data must have been an outlier or a fluke—a result that is unlikely to be replicated. Some students who follow this line of reasoning still employ “by chance” language. Notice how Anthony employed statistical terms as he reasoned 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.

Anthony interpreted the sampling distribution as evidence that babies are equally likely to choose the helper or the hinderer toy. In the observed sample, 14 out of 16 babies chose the helper, but because this value falls in the tail of the distribution, Anthony believed it was an unusual chance occurrence that should be disregarded.

At one point in her individual interview, Eva's reasoning about the helper vs. hinderer task 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, moments later, 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.
- Interviewer: 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 theory-based test. When it was pointed out that that she had not used the observed sample data yet, she remembered that the applet could be used to count samples as extreme as the observed statistic.

- Eva: (using the applet to count samples more extreme than 14 out of 16) That is so small.
- Interviewer: 3 out of 1000—what does that tell you?
- Eva: That getting 14 out of the 16 to choose the helper is very unlikely.
- Interviewer: 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 went on to contrast her new and previously held conceptions. Her explanations provided insight into the work of other students who conflated simulation and replication, making it possible for the researchers to link difficulties not previously seen as similar. Eva's trajectory suggests that it is possible for students to transition from one conception to another when confronted with contradictions in their reasoning. Teachers who recognize these difficulties can take appropriate steps to challenge their students' conceptions.

6. DISCUSSION AND IMPLICATIONS

In summary, although simulation-based inference methods offer a number of proposed advantages over theory-based inference alone, difficulties still arise. These difficulties 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. This section discusses how the findings of the present study are connected to existing literature and comments on the scope and implications of the findings.

6.1. COORDINATION OF LEVELS AND DIMENSIONS

Saldanha and Thompson (2002) described coordination of levels 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. (p. 264)

Disappointed with the results of their teaching experiment, Saldanha and Thompson decided that the simulation activities “were of such a complexity so as to essentially overshadow ideas of sampling variability” (p. 268). Although the present study also found student difficulty coordinating levels, students often resolved these issues and went on to reason about the statistical concepts under study, particularly when they were given statistical tools like applets or when they were asked follow-up questions by an interviewer or classmate. Thus, the current study may lead to a more optimistic appraisal of simulation-based inference.

Similarly, this study found that students’ coordination of real-world and hypothetical perspectives was “unstable.” 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 a description of students’ developing conceptual systems as “less like rigid and stable worlds than they are like ... shifting collections of tectonic plates” (Lesh & Doerr, 2003, p. 18).

In some cases, students were able to design a simulation and justify their choice of physical model from the hypothetical perspective, but later interpreted the simulated results from a real-world perspective. For example, in the incident above, Anthony treated the simulated distribution as a representation of real-world replication. However, earlier in the interview, he had chosen 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”—clearly stating his intention to represent the null hypothesis. 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, 2003).

The results of this study may also help explain 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, the present study confirms that the term “by chance alone” is not universally understood (Pfannkuch, 2005); in fact, students may incorporate this language into an alternative logic of inference. Second, students struggle to coordinate simultaneously 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.

These findings about student difficulties have implications for instruction, both for proactive instructional design and reactive responses to students. For example, teachers using a simulation-based approach may prioritize precise language and explicit statements of the assumptions. Aware of potential confusion between *number of samples* and *sample size*, they may decide to call simulated samples *trials* or *repetitions*, choosing a less ambiguous term and using it consistently. They may directly assess students’ ability to identify the observational units in a simulated distribution. They may ask students to predict where a simulated sampling distribution will be centered, thus encouraging them to acknowledge the assumption of the null hypothesis. There remains a need for future work that explores how to address these difficulties most effectively.

6.2. 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 numerous 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 the 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 should not be ignored.

6.3. TWO APPROACHES TO INFERENCE

This paper focuses on difficulties that arise in simulation-based inference, but the same conceptions may also exist in courses that employ theory-based 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 from multiple samples. Additionally, the logic of inference—which requires coordination of the real-world and hypothetical perspectives—remains unchanged across inferential approaches. In short, the results presented do not provide a basis for discarding simulation-based inference in favor of theory-based inference.

One salient result of the larger study is that simulation-based inference has the potential to make student thinking visible. Recall that students in this study were exposed to both simulation-based and theory-based inference in instruction. Using the tools made available in the course—including a problem-solving framework, a memorized list of conditions for inference, and a graphing calculator—many students were able to carry out theory-based significance tests quickly and efficiently by end of the year. However, when probed for details about the underpinnings of the theory-based approach in individual interviews, some revealed incomplete understanding of the logic of inference. Further, theory-based inference tasks, like those the students completed in preparation for the AP Statistics exam, did not generally lead to spontaneous discussion of statistical modeling or the logic of inference. Thus, these tasks did not provide as many opportunities for students to challenge each other’s conceptions or for a teacher/researcher to evaluate student thinking.

Of course, use of theory-based inference in instruction is not a monolithic pedagogical approach, and educators have been working to improve conceptual understanding from within this framework for decades. The current study has implications for that work, as it highlights conceptions that may go unnoticed in theory-based courses. For example, do students recognize theoretical probability distributions as sampling distributions of test statistics? Do they acknowledge the assumption of the null hypothesis? Beyond memorized rules, do they understand why sample size conditions are necessary for validity? Further research is necessary to explore how teachers that use theory-based approaches can discourage over-reliance on rules and procedures while encouraging engagement with these concepts.

Finally, some difficulties described in this article are specific to courses that employ *both* theory-based 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* theory-based tests. In these courses, empirically-derived sampling distributions are another set of models to represent outcomes under the null hypothesis. As described above, additional models can lead to confusion if students make inappropriate connections between theory-based and simulation-based approaches. However, these inappropriate connections must be weighed against the advantages of using simulations to complement theory-based inference. Though research sub-question (2) was not addressed in this paper, the larger study found that many students do make productive connections across simulation-based and theory-based inference methods. Other teacher-researchers have also reported successful use of simulation to improve understanding of theory-based methods (e.g., Lane-Getaz, 2017; Reaburn, 2014).

The current study does not provide a basis for comparing pedagogical approaches, because all students were exposed to both approaches in instruction, but it does have implications for teachers who choose to complement theory-based inference with simulations (as the authors of this study continue to do in their own practice). First, this study brings attention to the issue of combining inferential approaches. This issue may not surface if students are only presented with narrowly defined inference tasks—that is, tasks that clearly specify which model the student should use. Second, the findings highlight the importance of clear transitions from one inference approach to another. Providing students with a framework to make connections capitalizes on the inferential conceptions they already hold, and assessments that prompt students to reflect on these connections may lead to valuable insights for teachers, researchers, and the students themselves.

In recent years, statistics educators have demonstrated the potential of simulations to improve conceptual understanding of inference but 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.

REFERENCES

- Aquilonius, B. C., & Brenner, M. E. (2015). Students' reasoning about p -values. *Statistics Education Research Journal*, 14(2), 7–27.
[Online: [https://iase-web.org/documents/SERJ/SERJ14\(2\)_Aquilonius.pdf](https://iase-web.org/documents/SERJ/SERJ14(2)_Aquilonius.pdf)]
- 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, The 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 K. Makar, B. de Sousa & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9)*, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.
[Online: 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.), *Working cooperatively in statistics education. Proceedings of the Seventh International Conference on Teaching Statistics (ICOTS-7)*, Salvador, Brazil. Voorburg, The Netherlands: International Statistical Institute.
[Online: 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).
[Online: <https://escholarship.org/uc/item/6hb3k0nz>]
- 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*.
[Online: media.collegeboard.com/digitalServices/pdf/ap/ap-statistics-course-description.pdf]
- 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). Dordrecht, The Netherlands: Kluwer Academic Publishers.
- 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.
[Online: [https://www.stat.auckland.ac.nz/~iase/serj/SERJ6\(2\)_delMas.pdf](https://www.stat.auckland.ac.nz/~iase/serj/SERJ6(2)_delMas.pdf)]
- GAISE College Report American Statistical Association Revision Committee. (2016). *Guidelines for Assessment and Instruction in Statistics Education College Report 2016*.
[Online: <http://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—The International Journal on Mathematics Education*, 44(7), 883–898.

- Gould, R., Davis, G., Patel, R., & Esfandiari, M. (2010). Enhancing conceptual understanding with data driven labs. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS-8)*, Ljubljana, Slovenia. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://icots.info/icots/8/cd/pdfs/contributed/ICOTS8_C208_GOULD.pdf]
- Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research*, 7(1), 1–20.
- Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, 450 (557–559).
- 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.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS-8)*, Ljubljana, Slovenia. Voorburg, The Netherlands: International Statistical Institute.
[Online: 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.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS-8)*, Ljubljana, Slovenia. Voorburg, The Netherlands: International Statistical Institute.
[Online: 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* (pp. 3–34). Mahwah, NJ: Lawrence Erlbaum Associates.
- Knapp, H. R., & Fitzgerald, G. A. (1989). The antihypertensive effects of fish oil. *New England Journal of Medicine*, 320, 1037–1043.
- 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. D. (2005). Tinkerplots: Dynamic data exploration [Computer software]. Emeryville, California: Key Curriculum Press.
- Lane-Getaz, S. J. (2007). *Development and validation of a research-based assessment: Reasoning about p-values and statistical significance* (Doctoral dissertation). University of Minnesota.
[Online: www.stat.auckland.ac.nz/~iase/publications/dissertations/07.Lane-Getaz.Dissertation.pdf]
- Lane-Getaz, S. J. (2017). Is the *p*-value really dead? Assessing inference learning outcomes for social science students in an introductory statistics course. *Statistics Education Research Journal*, 17(1), 357–399.
[Online: [https://iase-web.org/documents/SERJ/SERJ16\(1\)_LaneGetaz.pdf](https://iase-web.org/documents/SERJ/SERJ16(1)_LaneGetaz.pdf)]
- Lehrer, R., & Schauble, L. (2007). Contrasting emerging conceptions of distribution in contexts of error and natural variation. In M. Lovett & P. Shah (Eds.), *Thinking with Data* (pp. 149–176). New York: Erlbaum.
- Lesh, R. A., & Doerr, H. M. (2003). 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 perspectives on mathematics problem-solving, learning, and teaching* (pp. 3–34). Mahwah, NJ: Erlbaum.
- Lock, R., Lock, P. F., Morgan Lock, K. L., Lock, E., & Lock, D. (2014). Intuitive introduction to the important ideas of inference. In K. Makar, B. de Sousa & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9)*, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://icots.info/icots/9/proceedings/pdfs/ICOTS9_4A3_LOCK.pdf]
- Malone, C., Gabrosek, J., Curtiss, P., & Race, M. (2010). Resequencing topics in an introductory applied statistics course. *The American Statistician*, 64(1), 52–58.

- 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 Practices & Council of Chief State School Officers. (2010). *Common Core State Standards*. Washington DC: Author.
- Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5(2), 241–301.
- 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: Springer.
- Reaburn, R. (2014). Introductory statistics course tertiary students' understanding of p -values. *Statistics Education Research Journal*, 13(1), 53–65.
[Online: [https://iase-web.org/documents/SERJ/SERJ13\(1\)_Reaburn.pdf](https://iase-web.org/documents/SERJ/SERJ13(1)_Reaburn.pdf)]
- 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.
- Starnes, D. S., Yates, D. S., & Moore, D. S. (2012). *The practice of statistics* (4th ed.). New York: W.H. Freeman.
- Tabor, J., Starnes, D. S., Yates, D. S., & Moore, D. S. (2012). *The practice of statistics: Annotated teacher's edition*. New York: W.H. Freeman
- Taylor, L., & Doehler, K. (2015). Reinforcing sampling distributions through a randomization-based activity for introducing ANOVA. *Journal of Statistics Education*, 23(3), 1–33.
[Online: <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.
[Online: [https://iase-web.org/documents/SERJ/SERJ11\(1\)_Tintle.pdf](https://iase-web.org/documents/SERJ/SERJ11(1)_Tintle.pdf)]
- Tintle, N., VanderStoep, J., Holmes, V.-L., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education*, 19(1).
[Online: <http://ww2.amstat.org/publications/jse/v19n1/tintle.pdf>]
- Wasserstein, R. L. & Lazar, N. A. (2016). The ASA's statement on p -values: Context, process, and purpose. *The American Statistician*, 70(2), 129–133.
- Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, 67(3), 223–248.
- 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.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9)*, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.
[Online: 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.
[Online: [https://iase-web.org/documents/SERJ/SERJ7\(2\)_Zieffler.pdf](https://iase-web.org/documents/SERJ/SERJ7(2)_Zieffler.pdf)]

CATHERINE CASE
Department of Statistics, University of Georgia
310 Herty Drive
Athens, GA 30602, USA