

INVESTIGATING NOVICE STATISTICS' STUDENTS REASONING ABOUT AND WITH SAMPLING DISTRIBUTIONS

by

CLAIRE MILLER

(Under the Direction of Amy B. Ellis)

ABSTRACT

Data are everywhere. Data collected from samples are often reported in the form of polls, medical studies, and advertisement information and an understanding of sampling distributions and statistical inference is important for evaluating data-based claims (Bargagliotti et al., 2020). Despite the importance of understanding statistical inference and sampling distributions, research shows that these ideas are challenging for students (Sotos et al., 2007). I argue that one source of difficulty is that the forms of reasoning that are expected and highlighted in statistics differ from those that are standard in mathematics. Mathematical reasoning is often deterministic, emphasizing proof and deduction. In contrast, statistical reasoning is about reasoning probabilistically, *not* deterministically. Thus, statistics involves a different type of reasoning from what is required in mathematics. Reasoning about data is about reasoning under uncertainty, a feature of both inductive and abductive reasoning. The purpose of this dissertation study was to understand how novice statistics students reason about and with sampling distributions, particularly from the perspective of Peirce's (1878) three classic forms of inferential reasoning—deduction, induction, and abduction.

Understanding how far sample outcomes vary from a population parameter is critical in reasoning about sample data. I found that students reasoned inductively when they generalized patterns that they observed in sample outcomes and, as a result, developing a better understanding of sampling variability. Statistical inference is often equated with proof by contradiction, a type of deductive proof. However, the conclusions drawn from sample data are *never* certain; thus, inference under uncertainty is *not* deductive logic. Instead, I found that reasoning abductively was particularly powerful for making inferences from sample data to a larger population. Students who reasoned abductively to repeatedly hypothesize multiple explanations for the sample data they observed provided a reasonable estimate for an unknown population parameter and coordinated multiple levels of distribution—a critical component in understanding sampling distributions. The findings of this study highlight the importance of developing research-based instructional practices that promote critical forms of reasoning, such as inductive and abductive reasoning, to support students in constructing important meanings and understandings of statistical concepts.

INDEX WORDS: Statistics education, sampling distribution, abductive reasoning, inductive reasoning, deductive reasoning

INVESTIGATING NOVICE STATISTICS' STUDENTS REASONING ABOUT AND WITH
SAMPLING DISTRIBUTIONS

by

CLAIRE MILLER

B.S., University of Georgia, 2008

M.Ed., University of Georgia, 2016

M.S., University of Georgia, 2024

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial
Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2024

© 2024

Claire Miller

All Rights Reserved

INVESTIGATING NOVICE STATISTICS STUDENTS' REASONING ABOUT AND WITH
SAMPLING DISTRIBUTIONS

by

CLAIRE MILLER

Major Professor:	Amy B. Ellis
Committee:	AnnaMarie Conner
	Christine Franklin
	Denise A. Spangler

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
December 2024

DEDICATION

I dedicate this dissertation to my grandfather, Carl O. Riggs, Jr., the smartest man I have ever known. He instilled in me a love and appreciation for mathematics as a young child.

Grandpa, I hope I have made you proud.

ACKNOWLEDGEMENTS

There are many people who supported me in this unique journey; without them, this dissertation would not be possible. I must first thank my advisor, Amy Ellis. Thank you for your unwavering confidence in me, even when I did not have confidence in myself. Thank you for your compassion, kindness, and support as I navigated this challenging and rewarding process. Thank you for pushing me to think deeply and for devoting countless hours of your time to meet with me and provide feedback on my work. Your mentorship and guidance have been invaluable to my growth as a researcher. To Anna Conner, thank you for giving me my first experience with research. Working on the CALC research team provided me with numerous opportunities to learn from you and other faculty across multiple disciplines. I am forever grateful for the opportunities to present at national and international conferences and co-author papers with you and other members of the CALC team. Most importantly, thank you for believing in me and convincing me to continue this journey when I felt like I had lost my way. To Denise Spangler and Chris Franklin, thank you for your time and continued support throughout my years in this program. Your expertise and feedback contributed greatly to my development as a statistics education researcher.

I am grateful to past and present faculty members in Mathematics Education at the University of Georgia for providing me with learning opportunities that advanced my thinking and understanding. Thank you to past and present graduate students for your camaraderie; only a small subset of people understand this experience and I could not have completed this degree without you as a support system. In particular, thank you to my cohort—Jenna, Mike, and

Anne—for your friendship and support. I am thankful to have had the opportunity to learn with and from each of you. Thank you to James, Anna, Dru, Aida, Lorraine, Erin, Ngutor, Mina, Shaffiq, Kelly, and Uyi for welcoming me back after my one-year hiatus; I never felt out of place, even after everyone in my cohort graduated before me. Our countless hallway and office chats made every day brighter. To Sarah, thank you for being my prospectus-writing buddy; achieving that long-awaited milestone together was so special. To Dru, thank you for always responding to my “Can you help me think through something?” texts; our conversations always pushed my thinking.

I am thankful for my friends and family for supporting me throughout my time in this program. To Mom and Bob, Linds, and the Foursome, thank you for being there from day one. No matter the distance between us, you have always been by my side. You are my permanent support system. I am grateful to the Athens Brunch Crew for cheering me on year after year and always making sure I celebrated the milestones in this program, however big or small. I could not have made it through these past few years without your continued friendship and support. And finally, to KCM, thank you for being my partner and my best friend. You supported me through the countless highs and lows and believed in me when I struggled to believe in myself. I could not have completed this journey without you, without US.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	v
LIST OF TABLES	ix
LIST OF FIGURES	x
CHAPTER	
1 INTRODUCTION AND PROBLEM STATEMENT	1
2 LITERATURE REVIEW	10
Why are Statistics Education and Sampling Distributions Important?.....	10
Challenges of Reasoning About and With Sampling Distributions.....	15
Research on Distribution.....	16
Research on Sampling Variability	23
Research on Sampling Distribution	28
What is Missing?.....	40
3 THEORETICAL FRAMEWORK	43
Reasoning in Mathematics and Statistics.....	43
Forms of Inferential Reasoning: Deduction, Induction, and Abduction.....	50
4 METHODS	62
The Clinical Interview Methodology	62
Participants.....	65
Study Design.....	66

Data Analysis	70
Summary	74
5 REASONING ABOUT SAMPLING DISTRIBUTIONS	75
Prior to Drawing Samples: Results from Task A.....	77
After Drawing Samples: Results from Tasks B and C	86
Summary	114
6 REASONING WITH SAMPLING DISTRIBUTIONS	120
Results from Task D	121
Summary	147
7 DISCUSSION	149
Reasoning About Sampling Distributions (Research Question 1).....	150
Reasoning With Sampling Distributions (Research Question 2).....	153
Limitations	157
Implications for Research and Teaching.....	160
Directions for Future Research	163
REFERENCES	166
APPENDICES	
A TASK A	181
B TASK B	182
C TASK C	185
D TASK D	187
E DESCRIPTION OF CLIPS IN INTERVIEW 1	189
F DESCRIPTION OF LORRAINE’S CLIPS IN INTERVIEW 2	190

LIST OF TABLES

	Page
Table 1: Hierarchies Characterizing Reasoning About Sampling Variability	25
Table 2: Description of Participants	66
Table 3: Data Collection Timeline.....	67
Table 4: Categories of Reasoning in Task A	80
Table 5: Categories of Reasoning in Tasks B and C	92
Table 6: Lorraine’s Inductive and Deductive Reasoning in Task B	95
Table 7: Comparing Predicted and Observed Outcomes	99
Table 8: A Visual Model of Students’ Shifts in Forms of Reasoning in Interview 1	116
Table 9: Students’ Reasoning Corresponding to Their Initial and Final Range of Values.....	118
Table 10: Categories of Reasoning in Task D	127
Table 11: Students’ Summaries in Interview 2	135

LIST OF FIGURES

	Page
Figure 1: A Pair of Perpendicular Line Segments	5
Figure 2: Three Triangles with Given Angle Measures.....	5
Figure 3: Two Points on a Coordinate Plane	6
Figure 4: Bakker and Gravemeijer’s Relation Between Data and Distribution.....	19
Figure 5: Using Results From a Least Squares Regression Analysis	46
Figure 6: Modeling a Pattern	48
Figure 7: Determining Whether a Point Lies on a Line.....	48
Figure 8: Theoretical Sampling Distribution for the North University Context.....	82
Figure 9: Box of Beads From Which Students Drew Samples in Task B	86
Figure 10: Parallel Dot Plots Displaying Predicted and Observed Outcomes	87
Figure 11: Screenshot of Web-Based Applet That Models the North University Context	88
Figure 12: Lorraine’s Predicted and Observed Sample Outcomes	96
Figure 13: Results of Eric’s 500 Random Samples	101
Figure 14: Results of Mindy’s 500 Random Samples	102
Figure 15: Tyra’s Predicted and Observed Sample Outcomes	111
Figure 16: Box of Beads From Which Students Drew One Sample in Task D	122
Figure 17: Testing a Prediction of 45% Using the Web-based Applet.....	123
Figure 18: The Result of Lorraine’s Simulation When Testing 60%	130
Figure 19: The Results of Corrina’s Simulations When Testing 40% and 60%	133

Figure 20: The Result of Julie’s Simulation When Testing 55%	138
Figure 21: The Results of Julie’s Simulations When Testing 52% and 60%	141
Figure 22: The Result of Julie’s Simulation When Testing 47%	141
Figure 23: The Result of Becky’s Simulation When Testing 30%	145
Figure 24: The Results of Waverly’s Simulations When Testing 65% and 45%	147

CHAPTER 1

INTRODUCTION AND PROBLEM STATEMENT

Understanding ideas of sampling, sampling distributions, and statistical inference are important for improving students' statistical literacy and productive citizenship (Franklin et al., 2007; Saldanha & Thompson, 2007), particularly given calls for students to “become critical consumers of statistically-based results reported in popular media” (GAISE College Report ASA Revision Committee, 2016, p. 8). Introducing students to the basic language and fundamental ideas of statistics will help them develop the statistical reasoning skills needed to evaluate evidence and claims based on data, enabling them to become better educated, well-informed members of society (Bargagliotti et al., 2020; Ben-Zvi & Garfield, 2004).

Although there is no singular definition for statistical reasoning that is widely used across statistics education research, researchers who have written about statistical reasoning (e.g., Ben-Zvi & Garfield, 2004; Garfield, 2002; National Council of Teachers of Mathematics, 2009) include similar statistical ideas, such as collecting and analyzing data, interpreting graphical displays and statistical summaries, and making inferences, all while attending to uncertainty and randomness. Furthermore, prominent policy documents related to statistics education (e.g., Bargagliotti et al., 2020; Franklin et al., 2007) emphasize these same statistical ideas in their descriptions of statistical reasoning and place the statistical problem-solving process at the forefront of statistical reasoning and thinking. This four-step process includes (1) formulating a statistical question that can be answered with data, (2) designing and carrying out a plan to appropriately collect data, (3) selecting and using appropriate methods to analyze the collected

data, and (4) interpreting the results of the analysis in relation to the original question (Bargagliotti et al., 2020; Franklin et al., 2007). Making sense of data and evaluating claims based on data necessarily requires an understanding of the “big picture,” including what question(s) the data answer, how the data were collected, how the data were analyzed, and if the interpretation of the results is appropriate.

Included within statistical reasoning and the four-step statistical problem-solving process is probabilistic reasoning (i.e., reasoning with uncertainty) and inferential reasoning. Inferential reasoning involves drawing conclusions from data, which is “a fundamental skill for participation in a global, technological, and data-driven society” (Noll & Hancock, 2015, p. 362), and necessarily involves a robust and productive understanding of sampling distributions. Therefore, understanding sampling distributions is a foundational concept for statistical reasoning more broadly.

Despite the importance of understanding statistical inference and sampling distributions, research suggests that ideas are challenging for students (Saldanha & Thompson, 2002, 2014; Sotos et al., 2007). Although many students may be able to carry out the calculations involved in formal inference procedures, such as confidence intervals and hypothesis tests, they often struggle to understand the underlying process and logic behind statistical inference (Chance et al., 2004). This difficulty stems from the complex and abstract concept of sampling distribution, which requires students to coordinate multiple ideas such as sample, population, distribution, variability, and repeated sampling (Chance et al., 2004; Kadijevich et al., 2008; Noll & Shaughnessy, 2012; Saldanha & Thompson, 2002, 2007).

In my experience as a high-school mathematics and statistics teacher, my students struggled to understand sampling distributions and how they relate to statistical inference,

regardless of my implementation of a variety of tasks and instructional strategies year after year. In addition, when teaching introductory statistics at the undergraduate level, I witnessed my students face these same difficulties. This robust issue—occurring across multiple populations, courses, and years—motivated me to explore how students understand sampling distributions. In this study, I make a distinction between reasoning *about* and reasoning *with* sampling distributions. When I discuss reasoning *about* sampling distributions, I refer to reasoning about the repeated sampling process involved in constructing a sampling distribution. When I discuss reasoning *with* sampling distributions, I refer to using sampling distributions to reason about the process of drawing conclusions from sample data. I believe that it is important to better help students develop productive ways of reasoning about and with sampling distributions. Moreover, the mathematics and statistics education research fields have not yet solved this problem.

Several researchers have investigated how students reason about distributions of data and sampling variability, but few have examined how students understand and reason about sampling distributions. The few studies that have investigated students' conceptions about repeated sampling and sampling distributions characterize students' perceptions mostly from a "part-whole" perspective, describing how students relate a sample to the population from which it was drawn. For example, Kahneman and Tversky (1982) distinguished between two perspectives when making decisions about the unusualness of a particular value of a sample statistic. A student reasoning with a *singular* perspective focuses only on the single sample drawn, whereas a student reasoning with a *distributional* perspective relates the individual sample to a collection of similar cases, for which probabilities can be estimated. Saldanha and Thompson (2002) also distinguished between two conceptions of sample, one involving part-whole relationships and the other involving multiplicative relationships. A student reasoning with an *additive conception of*

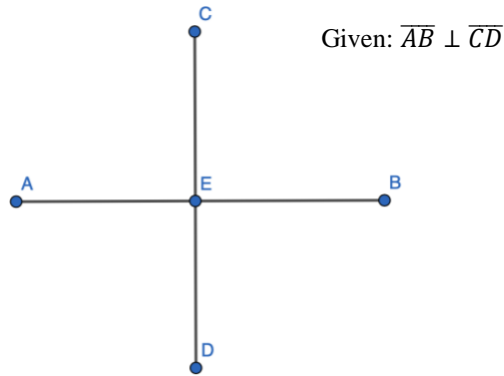
sample sees multiple samples as multiple subsets from the population that should resemble that population. In contrast, a *multiplicative conception of sample* (MCS) entails a quasi-proportional relationship between sample and population, in which multiple samples are seen as “multiple, scaled quasi mini-versions” of the population (Saldanha & Thompson, 2002, p. 267).

Although these different conceptions may help characterize the ways in which students describe a sample, there remain unanswered questions about how students understand and reason about and with sampling distributions. These conceptions do not address a bigger question about how the forms of reasoning in statistics differ from mathematical forms of reasoning. One promising route that differs from the extant research on conceptions of sample is considering how students understand and reason about and with sampling distributions from the perspective of forms of inferential reasoning, such as deductive, inductive, and abductive reasoning. I hypothesize that it can be productive to zoom out and investigate these broader forms of reasoning in statistics—and specifically with sampling distributions—that differ from what we know about how students reason mathematically.

I adopt the definitions of deductive, inductive, and abductive reasoning from Peirce (1878) and others (e.g., Reid & Knipping, 2010). *Deductive reasoning* is the process of drawing logically valid conclusions based on stated assumptions or rules and is thought to be the only form of reasoning that establishes a conclusion with certainty. For instance, consider a typical geometry example (see Figure 1). Suppose segments AB and CD are perpendicular and intersect at point E . All perpendicular lines intersect to form right angles. Therefore, I deduce (with certainty) that all four angles with vertex at point E are right angles.

Figure 1

A Pair of Perpendicular Line Segments



Inductive reasoning involves generalizing a rule from several observations. For example, suppose I measure the interior angles of three triangles of different shapes and sizes and notice that for each triangle, the sum of the interior angles is 180 degrees (see Figure 2). I induce that the sum of the interior angles for *all* triangles is (probably) 180 degrees. *Abductive reasoning* starts with an observation and involves hypothesizing what rule might lead to that particular observation. For example, suppose I observe points A and B on a Cartesian plane, as shown in Figure 3. I am not certain what function produced the two ordered pairs, but I might abduce that (possibly) a linear relationship produced the two ordered pairs.

Figure 2

Three Triangles with Given Angle Measures

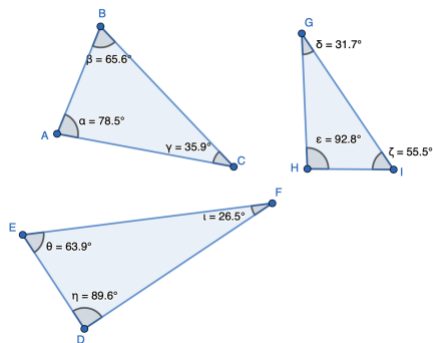
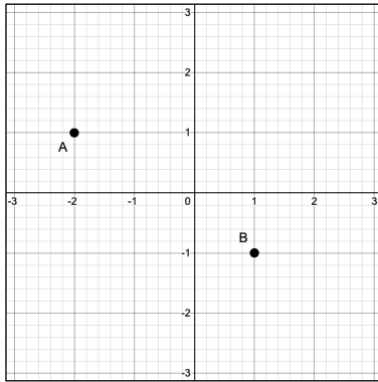


Figure 3

Two Points on a Coordinate Plane



Abrahamson (2012) explained,

Deduction, induction, and abduction are logical inference sequences, and each is modeled as a triadic structure leading from a premise through to a proposition whose truth is necessary (deduction), apparent (induction), or possible (abduction). All three types of inference are predicated on the epistemological framing that within any situation under inquiry there is some general *case* that necessarily implies a local *result* due to some *rule*...an individual person entering a situation may have only partial knowledge of the totality of these three notions. Which of these three notions the individual initially knows dictates the type of inference required in order to obtain more complete knowledge of the situation. (p. 633, emphasis in original)

In my experience teaching introductory statistics, some students seemed to reason about sampling distributions in a deterministic or deductive way, without attending to variability or uncertainty. For example, a student might reason that a particular sample should produce the same mean as the population from which it was drawn. That is, because the population has a particular mean, any sample drawn from that population should have the same mean. In contrast, suppose a student draws three samples from the same population and each of the three samples

produced equal sample means. This student might reason that, because all three samples produced equal sample means, the population from which they were drawn must also have the same mean. This student might be reasoning inductively because they are making a generalization about the population parameter after observing several samples with equal means. A student who might be reasoning abductively observes a sample with a particular mean and hypothesizes what population the sample was drawn from based on variability and probability.

Statistical reasoning involves reasoning with uncertainty and is fundamentally different from the deductive reasoning commonly seen in mathematics. Several researchers in mathematics education have studied deductive and inductive reasoning in the areas of reasoning and proof, but few have studied abductive reasoning. However, there is little to no work in the statistics education literature addressing these forms of reasoning. Kollosche (2021) noted, “How probabilistic and statistical reasoning relates to other forms of reasoning, how it justifies assertions, and how it contributes to our understanding of the world are not yet part of the academic reflections of the field” (p. 482).

In both mathematics and statistics, we see all three forms of reasoning. Conjectures and generalizations through inductive reasoning occur in both mathematics and statistics. Abduction occurs in both mathematics and statistics when hypotheses are constructed to explain some observation or event. Although mathematics and statistics share similar forms of reasoning, the emphases in each field differ. In mathematics there is an emphasis on deduction and proof, arriving at conclusions with certainty, and establishing truth. Although deduction can occur in statistics in the application of general rules to particular sets of data, such as using a common rule to determine outliers in a set of data, the emphasis in statistics is not on deduction. In statistics, the emphasis is on reasoning with uncertainty, a feature of both inductive and

abductive reasoning. Thus, it can be productive to understand how students reason about and with sampling distributions from this lens.

I hypothesized that reasoning abductively is a productive way to reason about and with sampling distributions, because abductive reasoning shares aspects of inferential reasoning. Reasoning abductively involves observing one or more cases and hypothesizing what rule might have produced those particular cases. Similarly, inferential reasoning involves observing one sample and hypothesizing what population would produce that particular sample. This is a promising avenue for statistics education research because characterizing the forms of inference students use when reasoning about sampling distributions provides insight into how students understand and make sense of sampling distributions. Moreover, this study can help mathematics education and statistics education researchers better understand the nature of other forms of reasoning beyond deduction and induction, such as abductive reasoning.

The purpose of this study was to understand how students reason about and with sampling distributions, particularly from the perspective of forms of inferential reasoning (i.e., deduction, induction, and abduction). My research questions for exploring this phenomenon were:

1. When given a population with a known parameter, what forms of reasoning do novice statistics students employ when determining what sample outcomes they expect and what sample outcomes they find surprising, and what do these forms of reasoning reveal about their reasoning *about* sampling distributions?
2. When given a population with an unknown parameter, what forms of reasoning do novice statistics students employ when making inferences from one sample outcome to the

population from which it was drawn, and what do these forms of reasoning reveal about their reasoning *with* sampling distributions?

CHAPTER 2

LITERATURE REVIEW

Why are Statistics Education and Sampling Distributions Important?

Statistics education is a fairly new field in the United States, particularly when compared to research conducted in the domains of algebra, geometry, trigonometry, and other common mathematics subjects. Although the push for the inclusion of statistics in school curriculum began in the early twentieth century, the importance of statistics education was not widely recognized until the 1980s due to the rise of a data-driven society. In addition, the emergence of personal computing and data analysis in industry and business motivated the need for statistics and data handling. Mathematics educators and mathematicians also argued for a revitalization of the curriculum: “Statistics and data handling [are] needed...if we [want] our citizens well-informed and our workers productive” (Scheaffer, 2006, p. 312). Following the rise of a data-driven society, statistics education continued to gain traction when major policy documents in education highlighted statistics content as a primary component of mathematics education. A brief timeline is provided below.

- 1983 – *A Nation at Risk* was published, which included probability and statistics as a major strand in the teaching of mathematics in high school.
- 1989 – The National Council of Teachers of Mathematics (NCTM) published the *Curriculum and Evaluation Standards for School Mathematics* and gave status to statistics education by making statistics and probability a content standard.

- 2000 – NCTM’s *Principles and Standards for School Mathematics* listed data analysis and probability as one of the five major content strands.
- 2005 – The *Guidelines for Assessment and Instruction in Statistics Education (GAISE)* College Report presented six recommendations for the teaching of introductory statistics and included a list of goals to promote the development of statistical literacy.
- 2007 – The *GAISE* pre-K-12 Report developed a framework that influenced state guidelines for mathematics, the development of teacher materials, and the Common Core State Standards in Mathematics (CCSS-M).
- 2010 – The CCSS-M were published and adopted by most states. The standards have a strong strand of statistics and probability in grades 6-12, but a weak presence in the elementary grades.
- 2015 – Precipitated by the adoption of the CCSS, the *Statistical Education of Teachers* (SET) Report distinguished between mathematical and statistical reasoning and outlined the content and conceptual understanding teachers need to know to help their students develop crucial statistical reasoning skills.
- 2016 – The *GAISE* College Report 2016 presented an updated list of learning goals for introductory statistics students and provided recommendations on statistical topics that might be omitted from or de-emphasized in an introductory course.
- 2020 – The *GAISE II* pre-K-12 Report incorporated the skills needed to make sense of new types of data relevant in today’s society, while maintaining the framework presented in the 2007 *GAISE* Report.

As evidenced by calls from important policy documents, the community of statistics and mathematics educators has come a long way in promoting the importance of statistics education

in the United States. Franklin et al. (2007) emphasized that “every high-school graduate should be able to use sound statistical reasoning to intelligently cope with the requirements of citizenship, employment, and family and be prepared for a healthy, happy, and productive life” (p. 1). Scheaffer (2003) echoed the need for students to develop statistical thinking to “engage effectively in quantitative situations arising in life and work” (p. 147).

Policy makers’ increased attention to the importance of statistics education is evident in the above timeline. Moreover, the concept of sampling distribution in particular has emerged as a key foundational construct within statistics education, not only for its importance in supporting an understanding of informal and formal statistical inference, but also due to the changing skills needed for living in a data-driven society (Garfield et al., 2015; Noll & Hancock, 2015; Noll, Redmond, et al., 2010; Saldanha & Thompson, 2007).

Societal Importance of Understanding Sampling Distributions

Data are everywhere. Data collected from samples are often reported in the form of polls, medical studies, and advertisement information. Citizens are often confronted with conflicting reports in the media, and an understanding of sampling distributions and statistical inference is important for evaluating data-based claims (Bargagliotti et al., 2020; Ben-Zvi & Garfield, 2004; Garfield et al., 2015; Saldanha & Thompson, 2007). Moreover, the ability to interpret reports from data and to evaluate data-based claims is a crucial skill needed for living in a society facing major issues, including effects from a global pandemic, climate change, economic highs and lows, and important social issues (Bargagliotti et al., 2020). Although all statistical reasoning is important, a coherent understanding of sampling distributions in particular is necessary to make appropriate interpretations of these data. For example, an article published in *Newsweek* reported on the results of a poll in the race for the Georgia Senate in 2022. The poll, conducted among

888 likely voters, indicated Raphael Warnock narrowly leading Herschel Walker 51% to 49% with a 3.2-point margin of error (Roche, 2022). What does this interval mean? How were these data collected? How was this interval calculated? Is this a large margin of error? How confident can we be in these estimates? To answer these questions, one needs an understanding of distribution, the stochastic nature of repeated sampling, sampling variability, and randomness. In short, one needs to be able to make sense of and reason about and with sampling distributions.

Classroom Importance of Understanding Sampling Distributions

The concept of sampling distribution is a foundational concept in statistical inference, which is the main focus of introductory statistics courses, whether at the secondary, undergraduate, or graduate level (Ben-Zvi et al., 2015; Lipson, 2003). Furthermore, sampling distribution is one of the most complex statistical concepts, because a coherent understanding requires the integration of several other important statistical ideas, such as distribution, the relationship between sample and population, variability in repeated sampling situations, randomness, and probability (Chance et al., 2004; Findley & Lyford, 2019; Heid et al., 2005; Kadijevich et al., 2008; Noll & Hancock, 2015; Noll & Shaughnessy, 2012; Saldanha & Thompson, 2002, 2007). These important statistical ideas leading up to and including sampling distributions and inference are woven throughout standards and policy documents for K-12 schooling.

According to the Common Core State Standards in Mathematics (CCSS-M), ideas of distribution and variability are seen as early as first grade in school mathematics (National Governors Association Center for Best Practices and Council of Chief State School Officers [NGACBP & CCSSO], 2010). However, the major ideas that contribute to an understanding of

sampling distributions are more formally introduced and explored beginning in the middle grades, through high school, and extending to introductory statistics courses at the college level.

Middle grades standards in statistics and probability address ideas of center and spread, with a focus on mean, median, interquartile range, and mean absolute deviation (NGACBP & CCSSO, 2010). Students in the middle grades are expected to describe distributions by their shape, center, and spread, and to display distributions with various graphical representations. According to the CCSS-M, students' first introduction to sampling and inference should occur in the middle grades, with the expectation that students will understand the relationship between a sample and the population from which it was drawn and use sample data to make inferences about a population.

Standards in statistics and probability at the high school level expand on these ideas to include more sophisticated summaries, representations, and interpretations of data. High school students are expected to choose appropriate measures of center and spread and to address extreme values when comparing distributions (NGACBP & CCSSO, 2010). In addition to understanding the process of making inferences about a population based on information gathered from a single sample from that population, high school students should investigate and interpret variability in repeated sampling situations by using simulation models to make conclusions about data.

The *GAISE College Report* included nine learning goals for students in college-level introductory statistics courses that focus on the development of statistical thinking (GAISE College Report ASA Revision Committee, 2016). Often, introductory statistics courses at the college level heavily emphasize statistical inference, whether informal or formal, for which sampling distributions are a foundation. Several of the goals outlined in the *GAISE College*

Report included major ideas that contribute to an understanding of sampling distributions (GAISE College Report ASA Revision Committee, 2016). According to the authors of the report, introductory statistics students should be able to summarize and interpret data both graphically and numerically, which entails being able to explain the significant role of randomness and variability in data. In addition to understanding sampling distributions, college introductory statistics students are expected to use sampling distributions to make both informal and formal statistical inferences.

Challenges of Reasoning About and With Sampling Distributions

Despite the importance of sampling distributions in statistics education, many researchers indicate that it is one of the most difficult statistical concepts for students to understand, due to its complex and abstract nature (Chance et al., 2004; Noll & Hancock, 2015; Noll, Redmond, et al., 2010; Noll & Shaughnessy, 2012), which contributes to the difficulties students experience when reasoning about and with sampling distributions. The concept of sampling distribution is one of the most complex statistical concepts because a coherent understanding requires the integration of several other important statistical ideas, including distribution, the relationship between a sample and the population, variability in repeated sampling situations, randomness, and probability (Chance et al., 2004; Findley & Lyford, 2019; Heid et al., 2005; Kadijevich et al., 2008; Noll & Hancock, 2015; Noll & Shaughnessy, 2012; Saldanha & Thompson, 2002, 2007). The concept of sampling distribution is one of the most abstract statistical concepts because one must imagine hypothetically drawing every possible sample of a particular size from a population with an unknown parameter of interest. Chance et al. (2004) described this challenge as, “a distinct, intangible thought process for most students” (p. 311). This thought process is also challenging for students because it requires an understanding of the relationship

between three different distributions—population distribution, data distribution, and sampling distribution—within the multi-level process of constructing a sampling distribution (Saldanha & Thompson, 2007).

When reasoning from a sample outcome to draw conclusions or make inferences about a population, students can often carry out the calculations involved in formal inference procedures, such as hypothesis tests and confidence intervals, but they have difficulty understanding the logic behind the underlying processes (Chance et al., 2004; Saldanha & Thompson, 2014). In addition, this process involves reasoning with uncertainty, which is difficult for students (Rubin et al., 1991). Because a coherent understanding of sampling distributions involves integrating several statistical concepts, understanding and reasoning about these concepts is crucial for reasoning about and with sampling distributions in productive ways.

Therefore, I begin my literature review by synthesizing the research on students' understanding of distribution and sampling variability, two of the major statistical ideas that undergird sampling distribution. Then, I review what researchers have found on students' understanding of sampling distributions, including reasoning about the process of constructing a sampling distribution and reasoning with sampling distributions to draw conclusions and make inferences from data. Within each of these sections, I also discuss the curricular and pedagogical interventions that have been investigated and report on the success of these strategies. Lastly, I point out what remains to be investigated with respect to students' understanding of sampling distributions, thus motivating the need for my study.

Research on Distribution

The concept of distribution in statistics is complicated, likely because it is used in various ways. In statistics, one might explore the *distribution of data*, the *probability distribution* for a

chance process, or the *sampling distribution* of a sample statistic, among others (Shaughnessy, 2007). In addition, the concept of distribution is seldom addressed as its own entity in school mathematics and statistics courses. Young students first encounter the notion of distribution when they construct pictograms to represent a favorite animal or a favorite food (Watson, 2009). Later, students construct more sophisticated representations of statistical data, such as dot plots and histograms, and learn to summarize features of distributions by describing the shape, center, and spread. In both of these experiences, rarely are students explicitly told they are learning about “distribution,” indicating the concept of distribution is treated as a “given” in statistical investigations (Wild, 2006).

Research on students’ reasoning about distribution emphasizes the importance of viewing data as an aggregate (Wild, 2006), or viewing data as one entity as opposed to individual observations. Rubin (2020) highlighted *data as aggregate* in his commentary of major themes in statistics education research, describing the dichotomy between a case-based view and an aggregate view of data. In reasoning about data, we rarely focus on the individual cases present in the data, but instead look for and interpret patterns in the data set as a whole. This distinction is important in understanding and reasoning about sampling distributions, as one needs to understand that a sample is simply one case out of a collection of similar cases.

We describe patterns by summarizing aspects of the data such as shape, center, and spread, attempting to distinguish the “signal” (center) from the “noise” (variability) (Bakker, 2004; Konold & Pollatsek, 2004; Wild & Pfannkuch, 1999). Shaughnessy (2007) noted, “data vary, samples vary, and distributions vary...variation occurs both within samples and distributions as well as across samples and distributions” (p. 972). The idea of distribution is based on patterns in variability. Although some readers may interpret *aggregate* and *distribution*

synonymously, Konold et al. (2015) argued aggregate is necessary but not sufficient in describing attributes, such as variability, of a distribution and called for an integration of multiple perspectives when working with data.

Wild (2006) described distribution as a structure or lens through which statisticians view variation in data. In comparison to statisticians who see data as an entity, or aggregate, novice statistics students tend to see data as a collection of individual observations (Bakker & Gravemeijer, 2004). Moreover, although many students can construct graphical displays of distributions of data, they do not generally use these representations to describe the data as a whole (Konold et al., 2015). Researchers have investigated the multiple perspectives students use when reasoning with data and distribution, exploring ways to support students in moving toward an aggregate view of data; I discuss these findings below.

Perspectives of Data and Distribution

In exploring how middle school students reason about data, Ben-Zvi (2004) identified two perspectives of data—local and global. A *local understanding* of data focuses on individual observations in a data set, whereas a *global understanding* involves perceiving the data as an entity and being able to identify and recognize patterns. In his study Ben-Zvi (2004) asked seventh-grade students to describe the trend in winning times for the men’s 100-meter Olympic race over time. He reported that students initially focused on the individual differences between observations, suggesting a local understanding of the data. Often, students view an individual observation as a characteristic of a particular person as opposed to the value of a variable (Ben-Zvi, 2004). For example, in the context of the 100-meter winning times, a student might view “12 seconds” as one runner’s personal winning time, rather than viewing the time as simply one value that the variable “winning time” takes for this set of data.

Bakker and Gravemeijer (2004) noted that a focus on individual observations is not uncommon among students in middle grades, even when calculating and interpreting measures of center and spread, such as mean and range, that are used to describe data distributions as a single entity. Bakker and Gravemeijer (2004) created a structure (see Figure 4) to represent the relation between data (as individual values) and distribution (as a conceptual entity). They examined aspects, such as center and spread, that are often used to describe both data and distribution.

Figure 4

Bakker and Gravemeijer's Relation Between Data and Distribution

distribution (conceptual entity)			
center mean, median, midrange, ...	spread range, standard deviation, inter- quartile range, ...	density (relative) frequency, majority, quartiles	skewness position majority of data
data (individual values)			

The structure can be read upward or downward, and experts in statistics can easily combine both perspectives. In contrast, novice students tend to view the structure from an upward perspective, viewing data as individual values for which they can calculate attributes such as mean, median, and range (Bakker & Gravemeijer, 2004). However, the ability to calculate these measures does not imply that students view them as attributes of the entire group (Konold et al., 2015). For example, students often use average as an adjective, as in “I am average height,” suggesting a focus on individual values. According to Prodromou and Pratt (2006), attending to measures of center and variability while viewing distribution as a collection of outcomes constitutes a *data-centric* perspective, or viewing data as an aggregate.

As discussed in the previous paragraph, research on students' understanding of data and distribution often emphasizes two general perspectives: a focus on individual observations and a focus on the entire group. Konold et al. (2015) expanded these broad characterizations when four distinct perspectives, based on different perceptual units, emerged from their analysis on students' reasoning with data. Students who perceived data as *pointers* did not refer to any apparent perceptual unit because they made no distinction between the data and the real world; the data were simply a reminder of the larger group from which the data were collected. In viewing data as *case values*, individual observations were the perceptual unit, because students with this view focused on characteristics of individual cases. Observations with similar characteristics were the perceptual unit when students perceived data as *classifiers*, in which students referred to a collection or a subset of the data that all shared the same characteristic. For the students who viewed data as *an aggregate*, the entire group was the perceptual unit, to which new groups of data could quickly be added.

Looking across these studies that address students' reasoning with data and distribution, the researchers highlighted the importance of viewing data as an aggregate. As I discuss next, researchers reported on curricular and pedagogical strategies to support students in developing a perspective of data and distribution that focus on viewing data as an entity. These interventions tend to include computer-based tools specifically designed for statistical analysis and data handling.

Curricular and Pedagogical Interventions

Some pedagogical interventions are aimed at helping students view data from a global perspective. For instance, when middle school students were tasked with describing patterns in winning times for the men's 100-meter Olympic race over time, Ben-Zvi (2004) noted that they

tended to focus on the individual difference between consecutive race times and struggled to let go of one race time that was an outlier. To support students in moving away from a local view and toward a more global view of the data, they were prompted to remove the outlier to broaden their focus on the rest of the data instead of narrowing their focus to one particular observation. In addition, after constructing a graphical representation of the data via computer software, students were asked to consider how changing the scales of the axes might affect the shape of the graph (Ben-Zvi, 2004).

Similarly, Bakker and Gravemeijer (2004) and Prodromou and Pratt (2006) engaged students in specially designed computerized statistical tools to support students in developing an aggregate view of data. Bakker and Gravemeijer (2004) used a series of web applets to encourage students to reorganize data in ways that were meaningful to them. Students were able to use the tools to visually represent the mean and shape of the distribution, both attributes that describe the data as a whole. Prodromou and Pratt (2006) conducted a design experiment in which middle school students changed parameters in a basketball-throwing activity in a computerized microworld. The microworld constructed graphical displays based on the outcome of each shot and related the parameters that the students changed to attributes in the distribution of data. Bakker and Gravemeijer (2004) flipped this idea on its head when they gave students particular attributes of data and asked them to construct graphical representations of the data, fostering a global view of data because students were unable to rely on individual data values. They also introduced the idea of “growing samples” to support students in developing an aggregate perspective of data and distribution. The idea of “growing samples” involves making predictions about characteristics of a distribution if more data were collected (Bakker, 2004). Lastly, students were asked about percentages of observations in a distribution to detract them

away from the mean and toward the entire distribution. This also gave students an opportunity to reason about relative frequencies, which involves comparing individual outcomes to the distribution of all outcomes, supporting a more global view of the data (Bakker & Gravemeijer, 2004).

The use of computerized tools can allow students to easily change parameters to highlight how these changes affect the distribution of data. For instance, Ben-Zvi (2004) found that directing students' focus away from an unusual observation and encouraging students to change the scale of the axes in graphical representations encouraged them to focus on global features of the graph, such as shape and direction, ultimately leading students to identify overall trends in the data. Bakker and Gravemeijer (2004) suggested that having students move back and forth between interpreting already made graphs and constructing unconventional graphs that are "meaningful and functional" (p. 155) to them helps students develop a more global view of data. Prodromou and Pratt (2006) echoed this by contending that, through interaction with the computerized microworld, students were able to pay attention to the emerging data and explain how features of the distribution related to the parameters they changed. These studies suggest that the use of computerized software and web-based applets can support students in moving away from a perspective of data that focuses on individual observations and toward a view of data as an aggregate.

Summary

Although students tend to focus on individual observations when considering data sets, researchers agree that a coordination of multiple perspectives is helpful in reasoning about data (Bakker & Gravemeijer, 2004; Ben-Zvi, 2004; Konold et al., 2015; Prodromou & Pratt, 2006). Researchers have found promising ways to support students in moving toward viewing data as a

single group. Asking particular questions about data that foreground data as an aggregate can support students in moving away from focusing on individual values (Bakker & Gravemeijer, 2004; Ben-Zvi, 2004; Konold et al., 2015). In addition, manipulating computerized displays and considering “growing samples” (Bakker, 2004; Bakker & Gravemeijer, 2004) can help students visualize trends and patterns in the entire distribution (Konold et al., 2015).

Research on Sampling Variability

The Lollies Task

In the early 2000s, there was a surge in research on students’ understanding of sampling variability due to responses on an open-ended item from the 1996 National Assessment of Education Progress (NAEP) that asked students to predict the number of red gumballs in a sample of ten taken from a gumball machine with a known proportion of red gumballs. In a random sample of several hundred responses to the gumball task, Zawojewski and Shaughnessy (2000) found that only one student provided a range of values rather than a single value.

Researchers and educators have reported that the difficulties students have in reasoning with uncertainty are likely due, at least in part, to the focus on correct answers in school mathematics coupled with the fact that questions of data and chance train students to report a single point answer, even when asked to estimate or make a prediction (e.g., estimate the number of successes, find the probability of an event) (Rubin et al., 1991; Shaughnessy, 2007).

Rubin et al. (1991) distinguished between two conceptions of sampling. *Sample representativeness* is the idea that a sample will often have similar characteristics as the population from which it was taken. In contrast, *sample variability* is the idea that multiple samples from the same population are not all the same as each other, nor are they the same as the population from which they were drawn. The results from the NAEP item suggested an

overreliance on sample representativeness with little or no attention to sample variability, motivating researchers to redesign the task (commonly referred to as the candy task in the United States or the lollies task in Australia) to explore how students think about variability in sampling situations. For example, Reading and Shaughnessy (2000) conducted task-based interviews with students ranging from ages 4 to 12 designed to compare students' responses on the lollies task across three different forms. Students were asked to (1) provide a list of values for each sample, (2) choose from seven multiple choice answers, and (3) give low and high bounds for the range of sample outcomes. Reading and Shaughnessy (2000) found that students seemed to be more capable of justifying their reasoning for center than for variability. Furthermore, they reported that students felt compelled to try to explain the "unexplainable" if the results from the physical simulation did not confirm their predictions; for example, students referred to variables such as hand size or how well the candies were mixed. Several researchers used variations of the lollies task to investigate students' reasoning about sampling variability, resulting in various frameworks and hierarchies to characterize this reasoning; I describe these frameworks in the next section.

Hierarchies Characterizing Reasoning About Sampling Variability

In this section, I report on three hierarchical frameworks that researchers developed to describe students' reasoning about variability in sampling situations. Each of these hierarchies emerged from research conducted with variations of the lollies task, among others. I summarize the frameworks in Table 1 and describe them in more detail below.

Torok and Watson (2000) investigated students' understanding of *isolated random variation* (in the context of variation in the lollies task) and *real-world variation* (in the context of variation in daily high temperatures). Based on analysis of interview data, they developed a

Table 1*Hierarchies Characterizing Reasoning About Sampling Variability*

Torok and Watson (2000)		Kelly and Watson (2002)		Shaughnessy et al. (2004)	
Level A: Weak appreciation of variation	Focus on individual outcomes with no consideration of variation	Level 1: Intuitive ikonik [<i>sic</i>] reasoning	Reasoning based on guessing, favorite numbers, size of hand, etc.; graphical representations lack ideas of frequency, clustering, or variation; no indication of proportional reasoning	Additive reasoning	Reasoning based on absolute frequencies with no consideration of variation
Level B: Isolated appreciation of aspects of variation and clustering	Acknowledge variation; basic understanding of clustering; weak proportional reasoning	Level 2: “More red” but consistent reasoning	Reasoning based on absolute frequencies; graphical representations include frequency but lack center	Proportional reasoning	Reasoning based on relative frequencies (i.e., proportions) with acknowledgement of variation
Level C: Inconsistent appreciation of variation and clustering	Acknowledge variation and clustering but have an inconsistent balance of the two; proportional reasoning when making predictions	Level 3: “More” or “half” red with centered reasoning	Reasoning with absolute or relative frequencies (i.e., proportional reasoning) and center; graphical representations indicate variation around the center	Distributional reasoning	Reasoning based on relative frequencies (i.e., proportions) and variation around the expected value
Level D: Good, consistent appreciation of variation and clustering	Appropriate balance of variation and clustering; strong proportional reasoning	Level 4: Distributional reasoning	Strong appreciation for proportion and variation; graphical representations indicate more variability than statistically appropriate, but include explanation of variation		

hierarchy consisting of four levels of developing conceptions of variation: weak, isolated, inconsistent, and good appreciation of variation and clustering (see Table 1). They found that one of the most salient differences between responses in the four levels was that students coded at lower levels of the hierarchy tended to make predictions with a single number, whereas students coded at higher levels tended to make predictions with an interval.

Kelly and Watson (2002) tested and refined the four levels in Torok and Watson's (2000) hierarchy. In their updated framework, Kelly and Watson (2002) included reasoning based on intuition in their description of the first level of reasoning (see Level 1 in Table 1). In their description of the fourth level of reasoning, *distributional* reasoning, Kelly and Watson (2002) included responses based on an explanation of variation (see Level 4 in Table 1). Consistent with Kahneman and Tversky's (1982) definition, students who reasoned distributionally indicated a strong appreciation for the proportion of red candies rather than absolute frequencies, or counts. This entails viewing one outcome as a particular instance of a group of similar outcomes, suggesting a perspective of data as an aggregate.

Drawing on their own previous work with the lollies task (see Reading & Shaughnessy, 2000), Reading and Shaughnessy (2004) first extended Kelly and Watson's (2002) hierarchy to include students' *description* of the variation and their explanation for the *cause* of the variation. Then, Shaughnessy et al. (2004) condensed this characterization of students' reasoning about sampling variability into three levels: additive, proportional, and distributional. In their variation of the lollies task, Shaughnessy et al. (2004) asked students to first predict the number of red candies in a sample of 10 from a bowl of 100 candies, of which 60 were red and 40 were yellow. Later, students were asked if they would expect the same number of reds every time if they were to take several samples of size 10. Students who provided justifications that relied on the

absolute frequency of red candies in the population, such as “there are more reds,” were categorized as reasoning *additively*. In contrast, students who justified their predictions based on the relative frequency of red candies in the population, for example, “the percent of red is higher,” were classified as reasoning *proportionally*. Expanding on Kelly and Watson’s (2002) definition of distributional reasoning, students who based predictions on relative frequencies *and* attended to both measures of center and variability when making predictions, responses such as “close to but not always 60% red,” were characterized as reasoning *distributionally*. Shaughnessy et al. (2004) found that most students relied on absolute frequencies rather than proportions in their responses, indicating they reasoned additively. Furthermore, they claimed that reasoning about variability proportionally is crucial in understanding sampling distributions and statistical inference.

Curricular and Pedagogical Interventions

In an attempt to support students in attending to variability in repeated sampling situations, several researchers reported on their use of physical samples. In many studies involving activities similar to the lollies task, researchers engaged students in making initial predictions, physically drawing samples from a bowl of candies, then comparing initial predictions to actual sample outcomes (see, e.g., Kelly & Watson, 2002). Even after physically drawing samples from a bowl of candies, many were hesitant to change their predictions, suggesting an overreliance on sample representativeness, or believing that the sample tells us *everything* about the population (Rubin et al., 1991). This was surprising because Shaughnessy et al. (1999) found evidence that physically taking samples increased the likelihood that a student would provide reasonable predictions in the lollies task.

Furthermore, Torok and Watson (2000) found that students tended to change their predictions after collecting samples of actual candies, suggesting that engaging students in physically taking samples might support them in attending to variability in repeated sampling environments. Although the results of these studies show varying levels of success with the use of physical sampling, several researchers agreed that providing students with more and earlier experiences attending to variability in sampling distributions could help students in balancing ideas of sample representativeness and sample variability (Reading & Shaughnessy, 2000; Rubin et al., 1991).

Summary

When making predictions about sample outcomes, students often gravitate toward providing point estimates over interval estimates (Torok & Watson, 2000). In addition, students tend to focus on absolute over relative frequencies in repeated sampling situations, suggesting a focus on individual observations rather than conceiving the sample outcome as one instance of a similar class of outcomes (Kahneman & Tversky, 1982; Kelly & Watson, 2002; Shaughnessy et al., 2004). Providing students with early and frequent experiences with sampling situations could support students in attending to sampling variability. Moreover, researchers reported the effectiveness of having students first make predictions, physically draw samples, then test their predictions against the actual sample outcomes (Rubin et al., 1991; Shaughnessy et al., 1999; Torok & Watson, 2000).

Research on Sampling Distribution

In my review of the research on students' understanding of sampling distributions, I distinguish between the research investigating students' understanding of the multi-level process involved in constructing a sampling distribution (what I refer to as reasoning *about* sampling

distributions) and the research examining students' understanding of using sampling distributions to draw conclusions from sample data to unknown populations (what I refer to as reasoning *with* sampling distributions). As I did in previous sections, I then discuss the curricular and pedagogical interventions that have been investigated with respect to sampling distributions.

Reasoning About Sampling Distributions

As noted earlier, part of the reason students experience difficulty when reasoning *about* sampling distributions is because it is an abstract concept that requires the coordination of multiple statistical ideas, including distribution and sampling variability. In addition, reasoning about sampling distributions requires an understanding of the relationship between three different distributions: (1) the (unknown, but hypothetical) population distribution, which models the distribution of a single variable across all members of the population; (2) the distribution of a single random sample, or the data distribution, which shows the distribution of a single variable across all members of a single random sample taken from the population; and (3) the sampling distribution, which models the variability in the sample statistic for *all possible* random samples taken from the unknown population. More specifically, reasoning about sampling distributions requires students to reason about the hypothetical behavior of all possible samples drawn from one unknown and hypothetical population, an extremely abstract thought process for many students (Chance et al., 2004; Harradine et al., 2011).

Recall, Konold et al. (2015) distinguished between a case value and an aggregate perspective of data and distribution. This dichotomy is also observed in students' perspectives of sampling distribution, as they tend to focus on individual samples rather than the collection of sample statistics (Saldanha & Thompson, 2002). According to Kahneman and Tversky (1982), when making decisions about the unusualness of a particular value of a sample statistic, a student

with a *singular* perspective focuses only on the single sample drawn, whereas a student with a *distributional* perspective relates the individual sample to a collection of similar cases, for which probabilities can be estimated. These perspectives were extended by Shaughnessy et al. (2004) to include additive, proportional, and distributional perspectives when reasoning about sampling variability, as discussed in an earlier section.

Building on the work of Shaughnessy et al. (2004), Noll and Shaughnessy (2012) investigated how students reason about sampling distributions by conducting teaching episodes with instruction on sampling distributions and variability with middle and high school students. After the teaching episodes, the students participated in clinical interviews. In this study, the students were first asked to make predictions and graphical representations of sample outcomes taken from a population with known parameters, then they were asked to reverse this process by making predictions about an unknown population based on sample outcomes. After initially applying Shaughnessy et al.'s (2004) framework to identify students' responses as additive, proportional, or distributional, Noll and Shaughnessy (2012) noticed that many of the responses indicated that students were in a transitional stage of their reasoning. Based on a more fine-grained analysis of the responses, they expanded the initial framework to include a transitional level between additive and proportional reasoning to include students' reliance on a single attribute in their thinking (i.e., shape, center, spread). When making predictions about sampling distributions, students in the research classes tended to reason at the additive or transitional levels, suggesting they were unable to integrate multiple aspects of the sampling distribution. When making predictions from sample outcomes to an unknown population, students who participated in the teaching episodes tended to reason at higher levels (proportional or distributional) than those in the comparison classes, although the researchers noted that students'

tendency to rely on centers was persistent in both groups. Similar to Torok and Watson's (2000) findings, Noll and Shaughnessy (2012) also found that students tended to give point estimates instead of interval estimates when predicting an unknown population parameter, again suggesting an overreliance on sample representativeness (Rubin et al., 1991).

Many researchers reported on deficiencies in students' understanding of statistical concepts such as distribution, variability, and sampling distribution. In contrast, Saldanha and Thompson (2002) and Findley and Lyford (2019) explored students' existing ways of thinking and reasoning about sampling distributions, and investigated how to support students in building on these ideas to develop a more coherent understanding of sampling distributions. For example, Findley and Lyford (2019) explored how introductory statistics students reason about sampling distributions and considered how students reconcile multiple and possibly contradicting ideas in their sense-making. Emerging from their analysis of interview data were ten "ideas or approaches to reasoning," which they called resources, that students used or might use when reasoning about sampling distributions: (1) repeated data values, (2) modeling likelihoods, (3) average relates to middle, (4) average relates to peak, (5) sampling distribution resembles population shape, (6) growing possibilities, (7) widening of range values, (8) stabilizing, (9) better representing population, and (10) becoming more accurate (p. 33). Although Findley and Lyford (2019) found that many of the identified resources corresponded to common misconceptions about sampling distributions and inference, they looked for what the students answered correctly and discussed how each of the resources could be used productively. For example, some students included in their reasoning that the sampling distribution resembles the population distribution. Even though this approach to reasoning is not correct for every sampling distribution, characteristics of the sampling distribution, such as shape, center, and spread, do

depend on the population distribution, which would be a productive application of this particular resource. Findley and Lyford (2019) recommended effective instruction to use students' existing ideas "as building blocks for conceptual models" and to offer alternative resources to elicit new perspectives from students (p. 40).

Saldanha and Thompson (2002) also positioned students as capable learners in their investigation of the development of students' understanding of sampling distributions. They engaged high school aged students in a teaching experiment designed to support students in developing a *distributional* perspective when reasoning about samples and sampling. Their aim was to help students conceive of sampling as "a scheme of interrelated ideas including repeated random selection, variability among sample statistics, and distribution" (p. 259). In general, students who performed better on the instructional activities developed a *multi-tiered scheme* of sampling, in which they were able to coordinate three different levels in the sampling process: (1) randomly select a sample of a given size from a single population and record a statistic of interest, (2) repeat the random selection process a large number of times and accumulate a collection of the statistic of interest for each sample, and (3) partition the collection of statistics based on a given threshold (p. 261). However, most students experienced difficulty in coordinating the three levels, which hindered their ability to imagine how a sample statistic might be distributed around the population parameter. Although most students understood that the value of the sample statistic varies from sample to sample, they were unable to extend the idea of sampling variability to the collection of sample statistics—the sampling distribution. Saldanha and Thompson (2002) conjectured these students held an *additive* conception of sample, one that focuses on part-whole relationships. From this perspective, multiple samples are seen as multiple subsets from the population. In contrast, a *multiplicative conception of sample*

(MCS) entails a quasi-proportional relationship between sample and population, in which multiple samples are seen as “multiple, scaled quasi mini-versions” of the population (p. 267). MCS balances ideas of sample variability and sample representativeness (Rubin et al., 1991), supporting students in developing a coherent understanding of sampling distributions.

Across the body of literature, these studies addressed the challenges students experience with the abstract nature of sampling distribution due to the many statistical ideas that sampling distribution entails. In addition, the studies referenced in this section addressed students’ difficulties understanding the distinction and relationship between the three different distributions and the multi-level process involved in the construction of a sampling distribution. Next, I discuss the literature on reasoning *with* sampling distributions, or the process of drawing conclusions and making inferences from sample data.

Reasoning With Sampling Distributions

The concept of sampling distribution is the underlying foundation for both informal and formal statistical inference: “Central to learning statistical inference is understanding that the variation of a given statistic (e.g., the mean) calculated from single random samples is described by a probability distribution—known as the sampling distribution of the statistic” (Harradine et al., 2011, p. 243). For this reason, researchers have investigated how students’ reasoning with statistical inference relates to students’ reasoning with sampling distributions

For example, as part of a larger study that investigated students’ abilities to relate ideas of sampling, variability, and sampling distribution (Saldanha, 2004), Saldanha and Thompson (2007) and Saldanha and McAllister (2014) explored students’ conceptual operations in making informal inferences based on empirical sampling distributions. In general, Saldanha and Thompson (2007) found that students were able to coordinate two levels of thinking: one level

attended to the collection of individual sample outcomes and the second level entailed partitioning the collection of sample outcomes based on a defined threshold. They hypothesized that the two-tiered structure of reasoning was important in students' development of a rule for deciding when two distributions were similar. However, students were surprised when their predictions about sample outcomes did not match the truth about the population. This motivated Saldanha and Thompson (2007) to design a second phase in which students were asked to compare multiple simulated sampling distributions from a small number of samples and determine if each distribution was unusual. Throughout instruction, students moved from focusing on a single sample outcome to attending to a collection of sample outcomes, reasoning *distributionally* about a collection of values of a sample statistic. Although the results of this study indicate that students can learn to make informal inferences based on their conception of a collection of values of a sample statistic, Saldanha and Thompson (2007) argued that developing ideas of repeated sampling and its aggregation is complex.

Building on the work of Saldanha and Thompson (2007), Saldanha and McAllister (2014) engaged students with simulation software intended to support them in making inferences from sample data. The lessons in this instructional sequence were designed to engage students in repeated processes in order to make inferences based on the results of a simulation. The results of the study indicated that a majority of students were able to use the distribution of the sample statistic to make informal inferences about the population. Common challenges students experienced included difficulty in understanding what the distribution of the sample statistic represented in the context of the problem situation, as well as difficulty in coordinating the multi-tiered repeated sampling process described by Saldanha and Thompson (2002).

Noll, Shaughnessy, et al. (2010) also used computer-simulated empirical sampling distributions in their investigation of students' statistical reasoning about distributions of data and sampling distributions. Students across various grade levels, ranging from middle grades to the graduate level, were asked to (1) compare four empirical sampling distributions constructed from a computer simulation, (2) infer which of the four was more likely to result from the given population with known parameters, and (3) decide which sampling distributions were "fake." Noll, Shaughnessy, et al. (2010) found that students across all grade levels attended to extreme values and shape when determining if a distribution was "real" or "fake," with little focus on sampling variability. In addition, students across multiple grade levels did not appear to consider the population parameter when making decisions based on extreme values in the distributions. Noll, Shaughnessy, et al. (2010) emphasized the difficulties students experienced with managing the tension between sample variability and sample representativeness, first noted by Rubin et al. (1991).

Curricular and Pedagogical Interventions

Not only do students experience difficulty reasoning about and with sampling distributions, but teachers also experience difficulty teaching students how to reason about and with sampling distributions (Garfield, 2002). One possible reason for this is that teaching statistical content at the K-12 level is left up to the mathematics teachers, and these mathematics teachers often have gaps in their knowledge of statistical content (Garfield, 2002; Shaughnessy et al., 2009). Another reason for this is because the reasoning process involved in constructing a sampling distribution and using a sampling distribution to make inferences from sample data is very complex and difficult to understand (Harradine et al., 2011). Garfield (2002) argued some teachers focus more on teaching the concepts and procedures rather than the reasoning process,

hoping students will develop the ability to reason about the multi-level process and logic behind making inferences as a result. In this section, I expand on the curricular and pedagogical interventions that have been attempted and investigated regarding reasoning about and with sampling distributions.

The sampling distribution is one of the most complex and abstract concepts in introductory statistics. Noll and Shaughnessy (2012) highlighted the need for earlier classroom experiences that focus students' attention on multiple attributes of a sampling distribution, especially variability, to support students in developing *distributional* reasoning. In addition, Noll, Shaughnessy, et al. (2010) called for extended studies to investigate how students construct knowledge about sampling distributions and statistical inference. Saldanha and Thompson (2002) proposed that designing instruction to target a *multiplicative conception of sample* and support students in conceiving sampling as a scheme of interrelated ideas might help students in developing distributional reasoning and a more coherent understanding of sampling and inference. Furthermore, Saldanha and Thompson (2007) and Saldanha and McAllister (2014) highlighted the need for teachers to help students to develop a conception of sampling distribution and statistical inference that entails coordinating multiple levels of reasoning. These strategies include incorporating opportunities for students to first make predictions about sampling outcomes, then engage in the repeated sampling process to construct a distribution of multiple sampling outcomes to help students build images of the multi-level repeated sampling scheme.

Other researchers have reported on the use of computer-based simulation to help students visualize abstract statistical concepts such as sampling distributions and inference (Noll & Shaughnessy, 2012; Noll, Shaughnessy, et al., 2010; Saldanha, 2004; Saldanha & Thompson,

2002, 2007, 2014), but few have directly addressed the effect of these simulation tools on students' understanding. However, Lipson (2002) and Garfield et al. (2012) explored the role of computer-based technology in the development of students' understanding of sampling distributions and statistical inference. Lipson (2002) investigated how using simulation software might support students in linking ideas of sampling distributions and statistical inference. She found that students often referred to the distribution of the sample instead of the sampling distribution when making inferences about the larger population, later identified by Chance et al. (2004) as a common misconception students hold when reasoning with sampling distributions. Although the use of the simulation software seemed to support students in clarifying some important aspects related to sampling distributions, Lipson (2002) noted that because all distributions are conditional (Wild, 2006), the simulation's specific role in illustrating the link between sampling distributions and statistical inference could not be established. Garfield et al. (2012) reported positive results from a three-month teaching experiment with a curriculum designed to use simulation- and randomization-based methods to support students in understanding statistical inference. Preliminary data from open-ended assessment items and observations of the course suggested students recognized the need to define and use a model to simulate outcomes from which they could make inferences. The authors claimed that students can be taught to reason statistically when making inferences by using statistical software designed for simulation- and randomization-based approaches to statistical inference.

Although some researchers suggested that simulation-based approaches are effective in helping students reason with sampling distributions and inference (e.g., delMas et al., 1999; Garfield & Ben-Zvi, 2007), others recommended caution when engaging students in using simulation (e.g., Hodgson & Burke, 2000; Watkins et al., 2014). Watkins et al. (2014), for

instance, found that using simulation to construct an approximate sampling distribution of the sample mean could introduce or reinforce misunderstandings about the relationship between characteristics of the sampling distribution and the population distribution as the sample size increases. For instance, at the end of a professional development course, high school teachers were asked to construct simulated sampling distributions for samples of size 5, 15, and 30. Although all nine teachers correctly explained the variability of the sample mean decreases as the sample size increases, most also explained that the mean of the simulated sampling distribution approaches the mean of the population distribution. Although this is the pattern the teachers observed in their simulated (and approximate) sampling distributions, the mean of the theoretical sampling distribution is equal to the mean of the population distribution. Watkins et al. (2014) provided strategies instructors might use to address the issue, such as using a very large number of samples in their simulation activities.

In his response to Watkins et al. (2014), Lane (2015) agreed with this “fix” and recommended that instructors make explicit that a simulated sampling distribution is only an approximation for the theoretical sampling distribution. Although simulations can provide dynamic visualizations of abstract concepts such as sampling distribution and statistical inference, students may not necessarily understand the relationships between the patterns they see (Chance et al., 2004). To improve the effectiveness of simulation software, Chance et al. (2004) and delMas et al. (1999) suggested having students make predictions about what the sampling distribution might look like and then check those predictions against sampling outcomes simulated with technology. In addition, Hodgson and Burke (2000) recommended using graphic organizers to direct students’ attention to salient features of the simulated sampling

distribution, and facilitating a debriefing session where students reflect on the underlying statistical theory behind the simulation.

Summary

When reasoning about sampling distributions, students tend to focus on the absolute frequencies of sample outcomes, indicating an additive perception of sampling (Noll & Shaughnessy, 2012; Saldanha & Thompson, 2002). In addition, the persistent focus on point estimates over interval estimates indicates a lack of attention to sampling variability when using sampling distributions to make inferences from sample data. Many researchers have called for instruction to address this issue by providing students with earlier and more frequent experiences with repeated sampling contexts (Noll & Shaughnessy, 2012; Saldanha & McAllister, 2014). More specifically, instruction should build on students' existing conceptions (Findley & Lyford, 2019), and support students in reasoning distributionally and developing a coherent understanding of the multi-level repeated sampling process (Saldanha & McAllister, 2014; Saldanha & Thompson, 2002, 2007). Furthermore, physical and computer-based simulations can be effective tools in helping students develop important understandings about sampling distributions and statistical inference (delMas et al., 1999; Garfield et al., 2012; Lane, 2015). However, care should be taken when planning and implementing instruction using simulation to avoid introducing or reinforcing misunderstandings (Watkins et al., 2014) and to support students in making connections between the patterns they see and the underlying statistical theory (Chance et al., 2004; Hodgson & Burke, 2000).

Often, standards described in policy documents begin with “understand.” For example, high school students are expected to “understand how sample statistics reflect the values of population parameters and use sampling distributions as a basis for informal inference” (National

Council of Teachers of Mathematics, 2000, p. 324). I agree with Saldanha and Thompson (2007) that these statements are problematic because it is unclear what it means to “understand” sampling distributions and statistical inference. Researchers tend to report on students’ misconceptions related to statistical concepts, such as distribution, variability, and sampling distribution. I argue that more research is needed that recognizes students as “capable reasoning agents” (Findley & Lyford, 2019, p. 26), and extends and refines existing models of students’ reasoning about and with sampling distributions.

What is Missing?

The literature discussed up to this point has chronicled the various perspectives students hold about statistical ideas, such as distribution, sampling variability, sampling distributions, and making inferences from sample data. According to Harradine et al. (2011), reasoning with sampling distributions to draw conclusions from sample data includes three main components: (a) the reasoning process, (b) the concepts, and (c) the computations. Research shows that although students can often carry out the calculations involved in formal inference procedures (i.e., significance tests and confidence intervals), their main difficulties in reasoning with sampling distributions lie within the first two components (Harradine et al., 2011). I have already discussed the research finding on students’ understanding of the concepts. Now I turn to what the literature says about the reasoning process.

Some researchers have criticized statistical inference for its flawed logical structure. For example, Falk and Greenbaum (1995) noted that the underlying process of formal hypothesis tests has been equated to proof by contradiction based on a *modus tollens* deductive argument, but that this logic does not hold. The logic of a hypothesis test is as follows: We observe a statistic (e.g., the mean) from a random sample that was drawn from some unknown population

in question. If the observed statistic from that sample significantly deviates from what is predicted by the null hypothesis, then we reject the null hypothesis in favor of the alternative hypothesis. In other words, if the conditional probability of obtaining a particular sample outcome given the null hypothesis, $P(\text{sample outcome} | H_0)$, is low, we reject the null hypothesis, and this corresponds to deducing a logical contradiction. However, Falk and Greenbaum (1995) claimed that the main idea in this logic is flawed. *Almost* a contradiction is not a contradiction: “The probabilistic counterpart of [the] logical deduction does not hold (p. 78).” In short, inference under uncertainty is *not* deductive logic.

According to Krueger and Heck (2017), making inferences from samples “honors the need for an inductive leap from the known (the sampled data) to the unknown (a hidden reality)” (p. 12). In referring to the use of these statistical inference procedures commonly used by researchers, Krueger and Heck (2017) stated that when researchers ask questions about unknown populations, they engage in inductive thinking as they think about what processes generated the sample data. I argue this is not inductive reasoning, but instead is abductive reasoning because this process involves hypothesizing what population produced the sample data. I argue that Harradine et al. (2011) also described statistical inference in an abductive way: “The process of assessing strength of evidence concerning whether or not a set of observations is consistent with a particular hypothesized mechanism that could have produced those observations” (p. 235).

Furthermore, Zieffler et al. (2008) defined reasoning with sampling distributions based on the notion of generalization: “Statistical inference can be defined as generalizing from a small sample to a larger population” (p. 41). Generalization is often referred to in discussions and descriptions of inductive reasoning, so it makes sense that if this is a common definition being used for statistical inference, then there exists a formal name of inference called *inductive*

statistical inference. However, making a generalization from only *one* sample to an unknown population does not match common definitions of inductive reasoning; a process that includes observing *multiple* cases, noticing some similarity or pattern across all cases, then defining a rule based on the pattern seen across those multiple cases. When I observe one random sample, this is simply *one* sample out of many *possible* samples. I do not observe many samples, find a pattern or similarity across those samples, then generalize this pattern to a larger group. Thus, reasoning from one sample to make an inference to a larger population, I argue, is not an inductive process, but an abductive process.

The extant research on students' understanding of sampling distributions identifies various conceptions students hold about sampling distributions, rather than how students reason about and with sampling distributions from the lens of forms of inferential reasoning: deduction, induction, and abduction. Although inductive reasoning is sometimes used in conjunction with using sample data to make inferences (see e.g., Groth, 2015; Krueger & Heck, 2017), I argue that drawing conclusions from data entails abductive reasoning. What is missing from the literature is a characterization of the forms of inferential reasoning (i.e., deductive, inductive, and abductive) students employ when reasoning about and with sampling distributions. Investigating students' reasoning through this lens provides insight into how students understand and make sense of this complex, abstract, and foundational concept in statistics. Furthermore, this research has implications for future research on the design of curricular and pedagogical interventions to support students in developing productive ways of reasoning about and with sampling distributions.

CHAPTER 3

THEORETICAL FRAMEWORK

Reasoning in Mathematics and Statistics

Definitions of reasoning include some process for arriving at a conclusion or making an inference. For example, Galotti (1989) proposed that reasoning is “mental activity that consists of transforming information (called the set of premises) in order to reach conclusions” (p. 333). Similarly, the National Council of Teachers of Mathematics (2009) defined reasoning as “the process of drawing conclusions on the basis of evidence or stated assumptions” (p. 4). Lithner (2000) defined reasoning as “the line of thought, the way of thinking, adopted to produce assertions and reach conclusions” (p. 166). Reasoning is important across all disciplines, and definitions of reasoning in various disciplines, such as mathematics and statistics, incorporate descriptions of mental activities and processes that are specific to the discipline.

According to the National Council of Teachers of Mathematics (2009), mathematical reasoning encompasses formal reasoning (e.g., proof), in which conclusions are logically deduced from stated assumptions, premises, definitions, and theorems. However, mathematical reasoning also includes investigation, exploration, conjecture, generalization, informal explanation, and justification (NCTM, 2009). More specific forms of mathematical reasoning include algebraic reasoning, proportional reasoning, and geometric reasoning, to name a few. In their attempt to conceptualize a model for mathematical reasoning, Jeannotte and Kieran (2017) conducted an analysis of research in the mathematics education literature on mathematical reasoning. They found four major elements of mathematical reasoning: (1) the activity/product

dichotomy, (2) the inferential nature, (3) the goal and functions, and (4) the structural and process aspect of mathematical reasoning (p. 6). The authors described the structural aspect as “the way in which the discursive elements combine in an ordered system that describes both the elements and their relation with each other” (p. 7). Jeannotte and Kieran (2017) cited philosopher Charles S. Peirce’s model of deduction, induction, and abduction, as well as philosopher Stephen E. Toulmin’s basic elements in his model of an argument (e.g., data, claim, and warrant) in their discussion of the structural aspect of mathematical reasoning. Conner et al. (2014) combined these two perspectives to examine the kinds of reasoning provided within collective argumentation in a high school mathematics classroom. Relevant to my study is Peirce’s model of deduction, induction, and abduction, which I will expand on in a later section.

Although statistics is considered part of the mathematical sciences and is incorporated into the mathematics curriculum in K-12 schooling, many statistics education researchers (e.g., Groth, 2015; Rossman et al., 2006; Scheaffer, 2006) argue that mathematics and statistics are two distinct disciplines. The “omnipresence of variability” (Cobb & Moore, 1997, p. 801) in data is what sets statistics apart from mathematics, and it is why statistical reasoning is inherently different from mathematical reasoning (Shaughnessy et al., 2009). Although statistics relies heavily on mathematics (e.g., probability theory), the two disciplines have different origins and foundational questions: mathematical reasoning involves deductive reasoning and proof, whereas statistical reasoning involves probability, randomness, and uncertainty (Groth, 2015). Rossman et al. (2006), in articulating some of these distinctions, gave a nice description of statistics as a mathematical science:

Statistics is a mathematical science...we use the singular “is” and not the plural “are” to emphasize that statistics is a field of study, not just a bunch of numbers. We use

“mathematical” as an adjective, because while statistics certainly makes use of much mathematics, it is a separate discipline and not a branch of mathematics. We use the noun “science,” for statistics is the science of gaining insight from data. (p. 323)

Mathematical models undergird statistical thinking and reasoning, but these mathematical models come with a degree of uncertainty, due to chance variability from random sampling or randomization of treatments in an experiment. Statistical reasoning connects ideas from data, random chance, and probability to make sense of, understand, and explain variability (Shaughnessy et al., 2009). Thus, statistics involves a different type of thinking and reasoning than mathematics (Cobb & Moore, 1997; Rossman et al., 2006). Moreover, the nature of discourse within each discipline differs. As Groth (2015) described it, “Much mathematical discourse is grounded in deductive reasoning and language of definitive proof, whereas statistical discourse is often characterized by inductive reasoning and qualified conclusions” (p. 4).

As mentioned in the introduction, we do see instances of different forms of reasoning (e.g., deductive, induction, and abductive) in both mathematical and statistical reasoning, but the emphases in each discipline differ. Although mathematical reasoning can include making predictions from messy data (e.g., mathematical modeling), these data are often not collected from a random process as they are in statistics. Thus, mathematical reasoning does not include reasoning with uncertainty due to chance variability. Instead, mathematical reasoning often includes deductive and inductive reasoning. In contrast, statistical reasoning often lacks definitiveness and involves probabilistic reasoning, reasoning with uncertainty, and inferential reasoning (Scheaffer, 2006). This is not to say that deduction never occurs in statistics, but the emphasis in statistics is drawing conclusions from data, which is a process that involves

probability and randomness. Furthermore, the conclusions and inferences made from data are *never* certain due to the nature of variability in data (Tran & Lee, 2015).

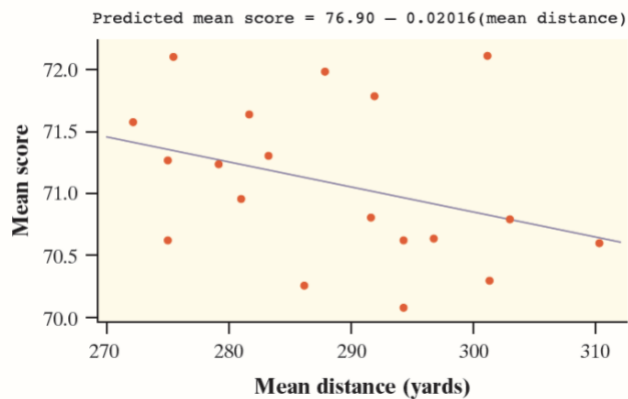
Consider three examples of problem situations that rely on the same mathematical idea: linear relationships. Figure 5 shows an example that involves using a least squares regression analysis to make predictions and inferences. This type of example is commonly seen in

Figure 5

Using Results From a Least Squares Regression Analysis

Do longer drives mean lower scores on the PGA Tour? Recent advances in technology have led to golf balls that fly farther, clubs that generate more speed at impact, and swings that have been perfected through computer video analysis. Moreover, today's professional golfers are fitter than ever. The net result is many more players who routinely hit drives traveling 300 yards or more. Does greater distance off the tee translate to better (lower) scores?

Data were collected on mean drive distance (in yards) and mean score per round from a simple random sample of 19 of the players on the Professional Golfers Association (PGA) Tour in a recent year. A scatterplot of the data and results from a least squares regression analysis are shown. The graph shows that there is a moderately weak negative linear relationship between mean drive distance and mean score for the 19 players in the sample.



- Predict the mean score per round for a PGA Tour player with a mean drive distance of 307 yards.
- Calculate the residual for the player with a mean drive distance of 275 yards and a mean score per round of 70.6.
- Do the data from this sample of 19 players provide convincing evidence of a negative linear relationship between mean drive distance and mean score per round for all PGA Tour players? Explain.

Note. This example was adapted from Starnes et al. (2014).

introductory statistics courses¹ and promotes multiple forms of reasoning. Notice in this example, the equation of the least squares regression line (LSRL) is provided. Both parts (a) and (b) involve deductive reasoning because both involve the application of a rule to a particular case; they simply require substituting values into equations to calculate the answer. However, interpreting the results of these calculations does require understanding that these data come from a random sample of PGA Tour players and that the linear relationship is not deterministic. The linear model does not provide the exact mean score per round for any PGA Tour player because these predictions are made with uncertainty due to chance variability. Moreover, the weak linear relationship seen in the scatterplot and quantified by the slope of the LSRL is only for this particular random sample of 19 PGA Tour players. If we were to record these same data from a *different* random sample of 19 PGA Tour players from that same year, would this weak linear relationship still hold? Part (c) addresses this question by asking students to make an inference about the population of *all* PGA Tour players², a process that involves *hypothesizing* what must be true about the population of all PGA Tour players to produce this particular sample. This process involves reasoning that is not deductive, but rather abductive.

The example in Figure 6 involves modeling a pattern with a linear function and the example in Figure 7 involves determining whether a specific point lies on a particular line. These types of examples are commonly seen in mathematics courses. Although both ask students to use linear relationships, the forms of reasoning that occur in each may differ.

¹ Although examples that resemble the one in Figure 5 might be used in mathematics courses, the emphasis is on estimating a line of best fit and using the equation to make predictions. The emphasis is *not* on how the data were collected, *not* on reasoning with variability due to chance, *nor* making inferences from the data to a larger group.

² To conduct a formal hypothesis test for the slope of the least squares regression line for the population of all PGA Tour players in this particular year, additional information is needed than what is currently provided in the example.

Figure 6

Modeling a Pattern

During preplanning, several mathematics teachers decided to cover the bulletin boards in their classrooms with fabric. The bulletins boards in the classrooms are various sizes, so the number of yards each teacher purchased differs. The table below shows the number of yards of fabric each teacher purchased and the price they paid.

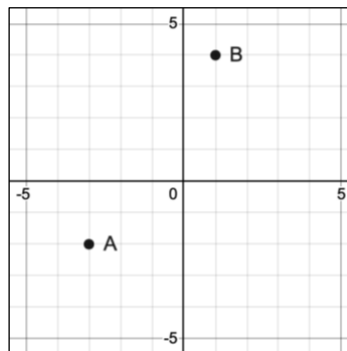
Number of yards of fabric	Price paid
7	\$45.50
10	\$65.00
5	\$32.50
2	\$13.00

- One of the mathematics teachers suggests covering the bulletin board in the hallway with fabric, too. Determine the cost of purchasing 8 yards of fabric.
- Another teacher measured incorrectly and needs 3 more yards of fabric. How much money will this teacher have to pay for these 3 yards?
- Write an equation that gives the price paid for any number of yards of fabric.

Figure 7

Determining Whether a Point Lies on a Line

The graph below shows two points, *A* and *B*. Use this information to answer the following questions.



Determine whether each of the following points lie on the line that passes through points *A* and *B*. Justify your reasoning.

- (0, 1)
- (13, 22)

The example in Figure 6 promotes both inductive and deductive reasoning, whereas the example in Figure 7 encourages deductive reasoning. In Figure 6, we might expect students to first examine the table of values for a pattern that holds for all values, using inductive reasoning. Next, students might apply their pattern, or rule, to the particular cases to determine the cost of the fabric for those specific numbers of yards, using deductive reasoning. In Figure 7, many students might use a general formula for a linear function (e.g., point-slope form) to determine the equation of the line that passes through points A and B . Substituting the slope between points A and B and one of the two points into the point-slope form of a linear equation is an example of applying a general rule to a particular case, or using deductive reasoning. Furthermore, determining whether the given points in parts (a) and (b) lie on the line also includes deductive reasoning in that students apply their rule (the linear equation) to a particular case (the given point), by substitution. On the one hand, students might determine a point is on the line because, when substituted into their linear equation, the equality holds. On the other hand, students might determine a point is *not* on the line because, when substituted into their linear equation, the equality does *not* hold. These are both forms of deductive reasoning, the former known as *modus ponens* and the latter known as *modus tollens*. I discuss these further in the next section.

As I already mentioned, the example in Figure 5 is commonly used in introductory statistics courses. At first glance, this example might also resemble some that are used in secondary mathematics classrooms, too. However, the standards relating to regression analysis in high school mathematics courses place importance on interpreting specific values in the linear function and performing calculations with the model (NGACBP & CCSO, 2010), processes that emphasize deductive reasoning. In contrast, when examples like this one are presented in introductory statistics courses, the goal is to use the regression model to make inferences about a

larger group, a process that emphasizes reasoning with uncertainty due to chance variability. The examples in Figure 6 and Figure 7 are commonly seen in the functions unit in middle grades or secondary mathematics classrooms (NGACBP & CCSSO, 2010) and do not involve statistical reasoning because (1) the values are not data that were produced from a random process, (2) there is a lack of variability due to chance, and (3) reasoning with uncertainty does not occur because the relationships are deterministic.

My goal in presenting and elaborating on the reasoning involved in these three examples is to illustrate that, although various forms of reasoning are used in both mathematics and statistics, the emphasis in mathematics is often on deductive reasoning, whereas the emphasis in statistics is often on inductive and abductive reasoning.

Forms of Inferential Reasoning: Deduction, Induction, and Abduction

As discussed in the previous section, most definitions of reasoning include some process of transforming given information to develop a conclusion or make an inference (e.g., Galotti, 1989; National Council of Teachers of Mathematics, 2009). Deduction, induction, and abduction are three classic forms of inferential reasoning and are the focus of this study. Peirce (1878) and others (Reid & Knipping, 2010) defined deduction, induction, and abduction in terms of a triadic structure involving a *case*, *rule*, and *result* that are linked in a particular order. A *case* is a specific observation that a condition holds. A condition describes an attribute of something, or a relation between things. For example, the statement “this bird is a swan” is a case, for which *being a swan* is the condition. A *rule* is a general proposition that states that if one condition occurs then another one will also occur. “All swans are white” is a rule that links the two conditions “being a swan” and “being white.” A *result* is a specific observation, similar to a case, but referring to a condition that depends on another one linked to it by a rule. “This swan is

white” is a result in this example. The order in which one links a case, rule, and result determines the kind of inference—deduction, induction, or abduction—needed to gain more knowledge about the situation. In deductive reasoning, a result is concluded from a rule and a case. In inductive reasoning, a case and a result lead to a rule. In abductive reasoning, a result and a rule infer a case. Consider the different ways case, result, and rule can be linked in the context of the swan example:

- *Deduction*: If I know that all swans are white (the rule) and that this bird is a swan (the case), then I can *deduce* (with certainty) that this bird is white (the result). I can deduce this without even seeing the bird.
- *Induction*: If I know that this bird is a swan (case) and that this bird is white (result), then I can *induce* that (probably) all swans are white (rule). This makes more sense if we talk about multiple swans (cases) that we’ve seen. Every swan I’ve ever seen is white, so I induce that all swans are white.
- *Abduction*: If I know that this bird is white (result) and that all swans are white (rule), I may *abduce* that (possibly) this bird is a swan (case). I see a bird that is white and all swans I’ve ever seen are white and I can’t really remember ever seeing a white bird that wasn’t a swan. So, I abduce that this bird is a swan.

In the remaining sections in this chapter, I first define and provide examples of deduction, induction, and abduction. Within each of these three sections, I identify and describe the various roles, functions, and types of deductive, inductive, and abductive reasoning. Because abduction is often confused with induction, I then compare and contrast abductive and inductive reasoning. Lastly, I argue why these three forms of inferential reasoning provide a promising lens to investigate how students reason about and with sampling distributions.

Deductive Reasoning

Deductive reasoning is commonly defined as a formal way of reasoning from the general to the particular that includes a sequence of propositions (see, e.g., Harel, 2014). The reasoner begins with given information (called premises) that they presume to be true, then the reasoner reaches necessary conclusions/inferences that are logically true and valid from the given set of premises (Stylianides & Stylianides, 2008). Deductive reasoning is the main focus in formal logic because it is the only kind of reasoning thought to establish a conclusion with certainty (Reid & Knipping, 2010). In deductive reasoning, a rule is applied to a case, yielding a result. In other words, deduction is the application of a general rule to a particular case (Peirce, 1878, p. 40). In the bird example above, the general rule “all swans are white” is applied to the particular case “this bird is a swan,” yielding the result “this bird is white.” This syllogism is of the type called *modus ponens*, Latin for “affirming the antecedent.” Modus ponens, a rule of inference used to draw logical conclusions, states that if some premise p is true and p implies q , then q is also true. The other form of deductive reasoning is *modus tollens*, Latin for “denying the consequent.” This rule of inference combines modus ponens with the contrapositive and states that if p implies q and q is false, then p is also false. Using the bird example, if “all swans are white” and “this bird is not white,” then “this bird is not a swan.”

Because deductive reasoning establishes certainty, it is often used for verification. However, verification is not the only role of deduction. Reid and Knipping (2010), for instance, identified other roles of deductive reasoning, such as explanation and exploration. Peirce (1878) argued that deductive reasoning cannot be used to explore new truths, because all necessary information for establishing a conclusion is already present in the premises. However, Reid and Knipping (2010) argued that “deductive reasoning can lead to the *experience* of discovering a

new truth” (p. 87) and gave an example of decoding a message: if you are given a coded message and the key to the code, then you possess the message and require no additional information to decode it. However, they go on to say that possessing the message is not the same as knowing what the message is. The claim that deductive reasoning does not create new knowledge represents an observer’s perspective, rather than the actor’s perspective, of the truth. In the context of decoding a message, the person who originally coded the message (i.e., the observer) possesses the “truth,” or knows the message, but has hidden the message from the decoder (i.e., the actor). I agree with Reid and Knipping (2010) that possessing the coded message and the key to the code does not mean that the actor knows or understands the message, but I argue that the process of decoding the message *does* create new knowledge for the actor because the actor now *knows* what the message says.

Consider a second example, in a statistics context. Suppose you are given a set of data and are asked to determine if the data contain outliers. Peirce (1878) would argue that you already possess the truth about whether or not the data contain outliers. Following their argument in the decoding a message example, Reid and Knipping (2010) might argue that, because you already know the rule for determining outliers, you require no additional information to determine whether any of the observations in the data are outliers. However, Reid and Knipping (2010) also suggested that “the act of deducing...can make what was implicitly known into something which is explicitly known, an experience not unlike learning something new” (p. 88). Based on this claim, possessing the rule to determine outliers is the not the same as knowing whether or not the data contain outliers. By applying the outlier rule to these particular data, or engaging in this deductive process, you created *new* knowledge because you now understand whether or not these data have outliers.

Although Reid and Knipping (2010) discussed that the role of deduction included verification, explanation, and exploration, Schurz (2017) zoomed out and argued that all forms of inference “have a justificatory (or inferential) and a strategic (or discovery) function, but to a different degree” (p. 153). The justificatory function refers to the justification of the conclusion based on the justification of the premises, whereas the strategic function refers to searching for the most promising conclusion for further testing (Schurz, 2017, p. 153). According to Schurz (2017), the justificatory function in deductive reasoning is foregrounded because the truth of the conclusion is guaranteed based on the truth of the premises. Deductive reasoning may include, however, a strategic function, because several different conclusions can be implied from the same premises.

Deductive Reasoning and Mathematical Proof. In mathematics, proof is often defined in terms of deductive reasoning (see, e.g., Stylianides & Stylianides, 2008) and some mathematics education researchers even claim that mathematical proof is synonymous with deductive reasoning (e.g., Harel & Weber, 2020). Both forms of deductive reasoning, *modus ponens* and *modus tollens*, are the foundation of several mathematical proof methods: direct proof and proof by mathematical induction are based on *modus ponens*, whereas indirect proof (e.g., proof by contradiction) is based on *modus tollens* (Stylianides & Stylianides, 2008, p. 108).

Although the role of proof in mathematics has traditionally emphasized verification, de Villiers (1990) criticized this one-sided view and argued that there are several other roles and functions of proof that are, in some cases, more important than verification. He expanded on Bell’s (1976) original distinction between proof as verification, illumination, and systematization to include five major roles and functions of proof that, he argued, are not mutually exclusive: (1) verification, (2) explanation, (3) systematization, (4) discovery, and (5) communication. In his

discussion, he compared these five roles within the areas of logical deduction, intuition, and quasi-empirical testing. Both Bell (1976) and de Villiers (1990) defined *verification* (justification or conviction) as addressing the truth of a statement, and although verifying or justifying for oneself is a different process than convincing someone else, these roles are similar enough that they are often discussed together (Reid & Knipping, 2010). In his discussion of the role of verification, de Villiers (1990) criticized a common belief about proof held by many mathematics teachers:

With very few exceptions, teachers of mathematics seem to believe that a proof for the mathematician provides absolute certainty and that it is therefore the absolute authority in the establishment of the validity of a conjecture. They seem to hold the naïve view...that behind each theorem in the mathematical literature there stands a sequence of logical transformations moving from hypothesis to conclusion, absolutely comprehensible, and irrefutably guaranteeing truth. However, this view is completely false. Proof is not necessarily a prerequisite for conviction—to the contrary, conviction is probably far more frequently a prerequisite for the finding of a proof. (p. 18)

When proof is used to provide insight into *why* a statement is true, the proof is being used as an *explanation*, or what Bell (1976) originally identified as *illumination*. De Villiers (1990) explained that mathematicians often seek a deductive explanation after reaching results that are either “intuitively self-evident” or “supported by convincing quasi-empirical evidence” (p. 20). Both Bell (1976) and de Villiers (1990) described the role of *systematization* as the organization of results into a deductive system of axioms, concepts, definitions, and theorems, and emphasized that intuition and quasi-empirical methods cannot ever fulfill this function. In stark contrast to the views of Reid and Knipping (2010) and Peirce (1878) that deductive reasoning

does not produce new knowledge, de Villiers (1990) asserted that there are many instances in which the function of logical deductive proof is *discovery* and exploration, and he criticized the naïve belief that deductive proof is used only for verification after new mathematical ideas are discovered through intuition and quasi-empirical methods:

There are numerous examples in the history of mathematics where new results were discovered/invented in a purely deductive manner; in fact, it is completely unlikely that some results (e.g., the non-Euclidean geometries) could ever have been chanced upon merely by intuition and/or only using quasi-empirical methods. Even within the context of such formal deductive processes as a priori axiomatization and defining, proof can frequently lead to new results. To the working mathematician proof is therefore not merely a means of a posteriori verification, but often also a means of exploration, analysis, discovery and invention. (p. 21)

De Villiers (1990) suggested that a possible reason for this common belief is due to how proof is often taught in schools, beginning with the result and ending with the proof. Lastly, proof can function as *communication*, a form of social interaction between mathematicians, teachers, and students, and a way to report and disseminate mathematical knowledge and results.

Inductive Reasoning

Recall, deductive reasoning can be characterized as the application of a general rule to a particular case or multiple cases. Inductive reasoning³ is often characterized as the reverse of deductive reasoning, or reasoning from particular cases to conclude a general rule (see e.g., Peirce, 1878; Reid & Knipping, 2010). As mentioned in the previous section, Reid and Knipping (2010) identified three main characteristics of deductive reasoning: (1) it is the application of a

³ Inductive reasoning does *not* refer to proof by mathematical induction, a type of deductive proof.

rule to a case to conclude a result, (2) it does not lead to new knowledge (though de Villiers (1990) and I disagree), and (3) it establishes conclusions with certainty. Opposing each of these in turn, Reid and Knipping (2010) characterized inductive reasoning in this way: (1) inductive reasoning concludes a rule from a case or several cases, (2) inductive reasoning uses existing knowledge to conclude new knowledge, and (3) inductive reasoning is probable, not certain. For Peirce (1878), “induction is where we generalize from a number of cases of which something is true, and infer that the same thing is true of a whole class. Or, where we find a certain thing to be true of a certain proportion of cases and infer that it is true of the same proportion of the whole class” (p. 472). In the psychology literature, Oaksford and Chater (2020) referred to this as *property induction* because one generalizes a property from one or more cases to new cases. In the bird example, the case “this bird is a swan” and the result “this bird is white” lead to the (uncertain) rule “all swans are white.” Or, more commonly, many similar cases associated with many similar results lead to a rule. So, we can extend the bird example to include more than one case: “I have seen several swans at various points in my life” and “all of those birds were white” lead to the rule “all swans are white.”

Although there do not seem to be commonly referred to roles and functions of inductive reasoning like what de Villiers (1990) identified and described for deductive reasoning, Reid and Knipping (2010) did define five types of inductive reasoning: (1) pattern observing, (2) predicting, (3) conjecturing, (4) generalizing, and (5) testing. Similarly, in their review of the literature on mathematical reasoning, Jeannotte and Kieran (2017) identified five mathematical reasoning processes that related to finding similarities and differences: (1) generalizing, (2) conjecturing, (3) identifying a pattern, (4) comparing, and (5) classifying. *Pattern observing* and *identifying a pattern* are similar in that they both involve noticing that several specific cases are

similar in some way, such as noticing that perfect square numbers alternate between odd and even numbers. *Predicting* is using known specific cases to make a claim about one or more unknown cases, such as making a claim that the next perfect square after 16 has the opposite parity. *Conjecturing* and *generalizing* both refer to making a general claim or rule based on specific cases, but Reid and Knipping (2010) distinguished these two types of inductive reasoning according to their need for additional verification. They argued that conjecturing requires additional verification, while generalizing does not. Similarly, Jeannotte and Kieran (2017) explained that conjecturing involves a search for regularity and is associated with the epistemic value of “probable” or “likely”, whereas generalizing expands from given cases to new cases and provides reason to believe the conclusion is valid. *Testing* is used to test predictions and conjectures. Just as testing is related to other types of inductive reasoning, *comparing* and *classifying* can take place during other reasoning processes, such as generalizing. It is important to note that when someone is reasoning inductively, they may use several or even all of these types of reasoning and processes. That is, one might first observe and identify a pattern by comparing similarities and differences across multiple cases, then predict what might come next in the pattern. They can test this prediction against the pattern and make a conjecture about the pattern. Testing the conjecture with several specific cases leads to a general rule, which is the outcome of inductive reasoning.

Abductive Reasoning

Charles S. Peirce is considered the first philosopher to define abduction, which he initially called *hypothesis* (Peirce, 1878). He described hypothesis as the inference of a *case* from a *rule* and a *result*. The case “this bird is a swan” is inferred from the supposed rule “all swans are white” and the result “this bird is white.” Peirce refined his original 1878 definition of

hypothesis and later renamed it abduction, reserving the word hypothesis for a conjecture that can be verified or refuted by facts (Peirce, 1960). However, he continued to characterize abduction in terms of rule, result, and case, emphasizing the logical form of abduction. Consider another comparison—and often-used example—of deduction, induction, and abduction based on the rule-case-result triadic structure:

- *Deduction*: If I know that all of the beans in the bag on the floor are white (the rule) and that a pile of beans on the table is from the bag on the floor (the case), then I can *deduce* (with certainty) that all of the beans in the pile on the table are white (the result).
- *Induction*: If I know that a pile of beans on the table is from the bag of beans on the floor (case) and that all of the beans in the pile on the table are white (result), then I can *induce* that (probably) all of the beans in the bag on the floor are white (rule).
- *Abduction*: If I know that all of the beans in the pile on the table are white (result) and all of the beans in the bag on the floor are white (rule), then I may *abduce* that (possibly) the pile of beans on the table is from the bag of beans on the floor (case).

Peirce also discussed the *explanatory* role of hypothesis and abduction: “Hypothesis is where we find some very curious circumstance, which would be explained by the supposition that it was a case of a certain general rule, and thereupon adopt that supposition” (Peirce, 1878, p. 472). His phrase “reasoning from effect to cause” in his description of hypothesis is yet another indicator of abduction as explanation (p. 477). Peirce emphasized the *exploratory/discovery* role of abduction with his example of Kepler’s discovery of the shape of the planets’ orbits:

The observed positions of the planets are P ,

If the shape of the orbits is an ellipse then the positions of the planets are P ;

Therefore, the shape of the orbits is an ellipse.

Kepler *abduced* and *discovered* a new rule that had not previously existed, now called Kepler's laws of planetary motion that describes the orbits of planets around the sun.

Although Peirce is credited as the first philosopher to define abduction in terms of its logical form and describe the roles of abduction, several others have characterized abduction as a form of inference and identified the goals of abduction in similar ways. Consider Magnani's (2009) definition of abduction:

Abduction is the process of inferring certain facts and/or laws and hypotheses that render some sentences plausible, that explain (and also sometimes discover) some (eventually new) phenomenon or observation; it is the process of reasoning in which explanatory hypotheses are formed and evaluated. (p. 8)

Schurz (2017) also agreed with the explanatory role of abduction by noting that abductions serve the goal of inferring the "unobserved *causes* or *explanatory reasons* of observed events" (p. 152). He also remarked that abductions can introduce *new* concepts, similar to the exploratory and discovery role identified by Peirce. Following Magnani (2001), Schurz (2017) referred to abductions that serve the purpose of explaining as *selective* abductions, because they serve the role of choosing the best explanation from several possible explanations⁴. In contrast, *creative* abductions serve the purpose of discovering or introducing new concepts, similar to how Peirce described Kepler's abductive discovery about the shape of the planets' orbits.

Comparing and Contrasting Inductive Reasoning and Abductive Reasoning

Induction and abduction are often confused because both are initiated by some observation and both involve a rule, whether that rule be inferred (induction) or hypothesized (abduction). Peirce (1878) distinguished the two: "[induction] infers the existence of phenomena

⁴ Formally termed the *inference to the best explanation* (IBE) by Harman (1965). See also Lipton (2004).

such as we have observed in cases which are similar...[abduction] supposes something of a different kind from what we have directly observed, and frequently something which it would be impossible to us to observe directly” (p. 480). Although both induction and abduction extend knowledge beyond what is observed, they each serve different purposes (Schurz, 2017).

Induction infers something about future observations. In contrast, abduction infers the explanation for or cause of an observation. Inductive reasoning is *generalizing* a rule from multiple observations that share some similar characteristic(s), whereas abductive reasoning is *hypothesizing* what rule could have produced a particular observation. In terms of evidence, inductive reasoning requires more evidence (i.e., multiple observations) to generalize a rule, whereas abductive reasoning requires minimal evidence (i.e., one surprising observation) to hypothesize a rule.

The aim of this study is to provide a characterization of how students reason about and with sampling distributions, from the lens of these three forms of inferential reasoning. I hypothesize that abduction is a more productive way to reason about and with sampling distributions than deduction and induction because constructing a sampling distribution and making an inference about characteristics of a population from characteristics of a single sample, I argue, includes the hypothetical, uncertain, and probabilistic process of abduction.

CHAPTER 4

METHODS

To better understand how students reason about and with sampling distributions, particularly from the perspective of forms of inferential reasoning (i.e., deduction, induction, and abduction), I used a series of task-based clinical interviews (Clement, 2000; Goldin, 2000) to investigate my research questions. Recall, the research questions I explored in this study are:

1. When given a population with a known parameter, what forms of reasoning do novice statistics students employ when determining what sample outcomes they expect and what sample outcomes they find surprising, and what do these forms of reasoning reveal about their reasoning *about* sampling distributions?
2. When given a population with an unknown parameter, what forms of reasoning do novice statistics students employ when making inferences from one sample outcome to the population from which it was drawn, and what do these forms of reasoning reveal about their reasoning *with* sampling distributions?

In the following sections, I first provide the theoretical background and characteristics of the clinical interview methodology to justify my use of the clinical interview. Next, I describe my participants and the data collection process by providing details on each phase of the study, including task details. Lastly, I describe my data analysis methods.

The Clinical Interview Methodology

The clinical interview method was originally developed by Piaget as an instrument for psychological research. Piaget was interested in the cognitive processes of children, and existing

methods (e.g., standard tests and observations) were insufficient for this purpose. For these reasons, he developed the clinical interview, a method of exploratory questioning with the goal of gaining insight into children's thought processes. This exploratory questioning is balanced by developing hypotheses about why the participant does what they do, which is accomplished by posing new tasks and asking more questions (Ginsburg, 1981, 1997).

Important characteristics of the task-based clinical interview include using open-ended tasks, asking questions that are often driven by the student's responses, developing and testing hypotheses, and building models of student thinking. In a task-based clinical interview, the participant is presented with a specific mathematical task that is open-ended, intended to give them the opportunity to expose their "natural inclination" (Ginsburg, 1981, p. 6). The design of the task and the sequencing of the interview should be based on a research question or an initial hypothesis (Cobb & Steffe, 1983; Goldin, 2000). In the clinical interview, presenting the participant with open-ended tasks promotes the nondeterministic nature of the interview process and gives the researcher the opportunity to be flexible in posing questions based on the participant's actions, maximizing the opportunity for discussion and reflection (Hunting, 1997).

Clarifying meaning is important as the researcher questions and poses tasks to students (Hunting, 1997). Although the researcher analyzes verbal and nonverbal behaviors and interactions, the researcher intends to facilitate rich verbalization that may provide insight into cognitive processes (Ginsburg, 1981; Goldin, 2000). Because the students' interpretation of the task is important for the researcher to understand, the researcher attempts to check and clarify ambiguous verbalizations. Thus, the role of language is central in clinical interview methods as interpretations are continually made by both the student and researcher (Cobb & Steffe, 1983).

Flexibility is another major feature of the task-based clinical interview (Ginsburg, 1981, 1997; Goldin, 2000; Swanson et al., 1981). Because the student and the researcher share ownership of the clinical interview environment, the direction the interaction takes is subject to change. Although the researcher sets the initial task and questions, subsequent decisions are contingent on the participant's lead (Ginsburg, 1997). The researcher makes decisions about questions and possible new tasks to pose based on the participant's responses. The flexible nature of task-based clinical interview methods allows the researcher the freedom to alter tasks to promote understanding, ask probing questions for further investigation of thinking processes, and pose new tasks on the spot to test hypotheses (Ginsburg, 1997). A clear research purpose guides the design and structure of the interview and may or may not include initial hypotheses about students' mathematical understanding of particular concepts.

Through the clinical interview process, researchers develop and test hypotheses about students' ways of thinking. These hypotheses are often made on the spot and are continually refined based on the student's responses to mathematical tasks and probing questions (Clement, 2000; diSessa, 2007; Ginsburg, 1981, 1997; Goldin, 2000). The task-based clinical interview is an opportunity to gather evidence and construct a model of students' current mathematical knowledge (Hunting, 1997). In using clinical interview methods, researchers continually formulate and test hypotheses about why the participant displays particular behaviors and verbalizations. This process aids researchers in constructing models of students' cognitive structures and processes (Clement, 2000).

The goal of the clinical interview method in mathematics education research is to gain insight into participants' *natural and existing* ways of thinking (Clement, 2000; diSessa, 2007; Ginsburg, 1981, 1997). Because I wanted to understand students' existing ways of thinking and

reasoning about and with sampling distributions, task-based clinical interviews were appropriate to investigate my research questions.

Participants

Many undergraduate programs across multiple universities and colleges require students to take an introductory statistics course. It is unsurprising that several undergraduate degree programs, such as those in STEM-related fields, would require additional courses in statistics. However, many of the undergraduate degree programs with a requirement of introductory statistics, such as fashion merchandising, might be surprising. An undergraduate degree in fashion merchandising, among others, has no additional requirements for statistics courses. Those students who go on to take more advanced statistics in their undergraduate or graduate coursework have additional opportunities to develop their reasoning about and with sampling distributions. However, for the students for whom introductory statistics is a terminal course, there are no additional classroom opportunities to develop their reasoning about and with sampling distributions. As mathematics and statistics education researchers, we need to find ways to support *novice* statistics students' understanding of sampling distributions now, so that they can develop the skills and reasoning necessary to be informed citizens with the ability to assess claims from data presented to them in their everyday lives. Before we can find ways to support novice statistics students in developing productive ways of reasoning about and with sampling distributions, we must first understand *how* these students are currently reasoning about and with sampling distributions.

For these reasons, I recruited 11 introductory statistics students at a large public university in the southeastern United States to participate in my study. All students had recently completed an introductory statistics course in Fall 2022 or Spring 2023, which was required for

their major. In Table 2, I identified each participant with a pseudonym, the most recent year they completed in their undergraduate studies, and their major. The recently completed introductory statistics course was a terminal course for all students except for Tami and Eric. The introductory statistics course from which I recruited participants was a large, primarily lecture-based course in which six sections were offered each semester, with 180 available seats for each section. The course content included common introductory statistical topics, such as experimental design, probability, sampling distributions, inference, and linear regression. In addition, there was a lab component taught by a teaching assistant. In the weekly lab sessions, students were introduced to a web-based statistical software package for analyzing data and used it to explore the big ideas covered in the lecture sessions that week.

Table 2

Description of Participants

Participant	Most Recent Year Completed in Undergraduate Studies	Major
Tami	Third	Computer Science
Julie	Second	Sociology; Psychology
Jess	Third	Pharmaceutical Sciences
Becky	Second	Entertainment Media (intended)
Lorraine	First	Psychology
Tyra	First	Exercise and Sports Science
Waverly	Fourth	Geography
Eric	Third	Computer Science
Lyla	First	Biology; Psychology
Mindy	Third	History
Corrina	First	Psychology

Study Design

The design of my study included two sets of task-based clinical interviews, one investigating students' reasoning *about* sampling distributions (Research Question 1), and one

investigating students' reasoning *with* sampling distributions (Research Question 2). In Interview 1, I engaged each student in a series of three tasks. The aim of this interview was to investigate how students reasoned under uncertainty when given a population with a known parameter. In Interview 2, I engaged the same students in a fourth task. The aim of this interview was to investigate how students reasoned under uncertainty when given a population with an unknown parameter. Specifically, I aimed to examine the forms of reasoning students employed when drawing conclusions from one sample outcome to a larger population. The time between each interview varied by participant, as shown in the data collection timeline provided in Table 3. All interviews lasted at most 75 minutes and were video and audio recorded using two cameras; one camera faced the student, capturing their facial expressions, gestures, and explanations; the second camera captured students' work from overhead. In addition, student's use of a web-based simulation applet was recorded in real time using a screen recording function on the tablet or via laptop connection.

Table 3

Data Collection Timeline

Participant	Interview 1 Date	Interview 2 Date
Tami	June 27, 2023	July 06, 2023
Julie	June 28, 2023	July 07, 2023
Jess	June 29, 2023	July 20, 2023
Becky	July 06, 2023	July 19, 2023
Lorraine	July 08, 2023	July 22, 2023
Tyra	July 18, 2023	July 19, 2023
Waverly	July 22, 2023	July 30, 2023
Eric	July 31, 2023	July 31, 2023
Lyla	August 13, 2023	August 27, 2023
Mindy	August 22, 2023	September 05, 2023
Corrina	August 23, 2023	August 28, 2023

Task Descriptions

Drawing from the literature base on students' understanding of sampling distribution and my experience with teaching this concept to statistics students, I identified critical aspects of understanding sampling distributions, including balancing ideas of sample representativeness and sample variability (Rubin et al., 1991), viewing data as an aggregate (Konold et al., 2015), and coordinating the multiple levels involved in the repeated sampling process (Saldanha & Thompson, 2002). To explore how students reasoned in a repeated sampling environment with a known population parameter, I designed a series of tasks that addressed these critical aspects of understanding sampling distributions.

Interview 1 Tasks. The three tasks I designed for Interview 1 presented students with multiple opportunities to reason about sample outcomes from a population with a known parameter. I created Task A (see APPENDIX A) to understand how students initially reasoned about and understood variability in a sampling situation, prior to drawing samples. I aimed to understand how students balanced ideas of sample representativeness and sample variability (Rubin et al., 1991). In Task A, I asked students to predict the outcome of one random sample drawn from the population of interest and compare this outcome to a friend's hypothetical outcome. Additionally, students determined what sample outcomes would surprise them. To conclude this task, students provided an interval estimate for the outcome of a single random sample drawn from the population of interest.

In Task B (see APPENDIX B), students used a physical simulation tool to draw samples; the tool was designed to model the population of interest and allow students to engage in a repeated sampling process to further explore sampling variability. I designed Task B to gain further insight into students' reasoning about sampling variability in a repeated sampling

environment by asking students to make predictions and comparisons. Several researchers reported the effectiveness of having students make a prediction about sample outcomes and then compare their prediction to sample outcomes after physically drawing samples (Chance et al., 2004; Shaughnessy et al., 1999; Torok & Watson, 2000). Furthermore, students' comparisons of their predicted and observed outcomes provide insight into how they view a collection of sample outcomes, either as individual outcomes or as an aggregate (Konold et al., 2015). In Task B, students made predictions about sample outcomes, physically drew samples with the simulation tool, compared their predictions with their actual results, and explained their thought process through the two sampling prompts.

Students used a web-based applet (Rossman & Chance, n.d.) to simulate taking a very large number of samples in Task C (see APPENDIX C). In this task, I asked participants to interpret parameters in the applet and explain outcomes based on the simulation applet to gain insight into how they interpreted and reasoned about the multi-tiered repeated sampling process (Saldanha & Thompson, 2002). Students' examination of a very large number of sample outcomes provided further insight into how they viewed a collection of sample outcomes (e.g., case-values, aggregate). In addition, students determined surprising outcomes and provide an updated interval estimate for the outcome of a single random sample drawn from the population of interest.

Interview 2 Task. I designed Task D (see APPENDIX D) to investigate how students reasoned *with* sampling distributions by asking them to make an inference about a population with an unknown parameter. In Task D, students initially used a physical simulation tool to draw one random sample; the tool was designed to model the population of interest. Throughout this task, students made and tested many predictions for the unknown population parameter from the

result of their single random sample. By continuing to ask students how the results of their simulation related to their initial sample outcome, I gained insight into their understanding about the relationship between sample and population. In addition, by asking students to explain how they could use simulation to gather evidence for their prediction, I examined students' ways of reasoning when drawing conclusion about the unknown population parameter.

Although these four tasks were well-thought out and I designed them to gain insight into the forms of inferential reasoning novice statistics students employed when reasoning about and with sampling distributions, there were times during the interviews when I made changes to the protocol based on my interactions with the student. These changes included asking additional follow-up questions or posing additional prompts. During the interview process, I developed initial hypotheses about students' ways of reasoning about and with sampling distributions and continued to refine and test these hypotheses by asking additional probing questions on the spot (Clement, 2000; diSessa, 2007; Ginsburg, 1981, 1997; Goldin, 2000) or by posing additional tasks. These changes were necessary to test and refine my hypotheses and provided me with the opportunity to gather additional evidence to construct a model of students' understanding (Clement, 2000).

Data Analysis

I used Transana (Woods, 2023) qualitative data analysis software to organize, transcribe, and analyze my data. My analytical process included retrospective analysis (Simon, 2019; Steffe & Thompson, 2000) and both convergent and generative approaches (Clement, 2000). In this section, I first provide an overview of how I used these three approaches in my analysis. Then I describe my analysis in more detail, organized by my adaptation of Simon's (2019) three-level approach to analyzing qualitative data .

Overview

Although retrospective analysis is often used in conjunction with ongoing analysis in the teaching experiment methodology (see e.g., Steffe & Thompson, 2000), it is also a useful approach to data analysis in other methodologies because it is a method for both analyzing data after the data collection phase has been completed and “developing models of student thinking” (Simon, 2019, p. 115). My primary goal of conducting retrospective analysis was to understand and construct models of student’s reasoning about and with sampling distributions.

Clement (2000) explained that clinical interviews can range from *generative* approaches to more *convergent* approaches. *Generative* clinical interview studies have more of an exploratory nature, with the goal of generating new theoretical models of mental structures and processes that explain new phenomena. In contrast, *convergent* approaches use previously developed models to document instances of relatively familiar phenomena. In this study, I aimed to characterize novice statistics students’ understanding of sampling distributions by examining the forms of inferential reasoning (e.g., deductive, inductive, abductive) they employed. I used Peirce’s (1878) case, rule, and result to guide my analysis, constituting a more convergent approach. However, characterizing students’ reasoning is much more complex than simply labeling their reasoning as deductive, inductive, or abductive. Although my study focused on these three forms of inferential reasoning, these broader forms of reasoning did not provide a complete model of students’ understanding of sampling distribution. By examining patterns of reasoning of individual students and across the participant pool, I developed emergent themes to describe students’ reasoning more broadly, constituting a more generative approach. These emergent themes highlighted important aspects of students’ understanding of sampling distribution beyond the forms of reasoning they used in each segment of data.

Levels of Analysis

My analysis process began with transcribing interview data from both sets of interviews. As I transcribed students' verbal utterances, I included descriptions of their gestures and screenshots of their written responses and their use of the web-based applet. This process allowed me to familiarize myself with the data and engage in initial exploration of students' reasoning. Next, I adapted Simon's (2019) three-level approach to analyzing qualitative data. In the first level, I analyzed the interview transcripts line-by-line and identified segments of data in which participants provided an explanation or a justification. Each of these segments, which I called reasoning clips, corresponded to at least one prompt in a task. Although both interviews were semi-structured, Interview 1 was more structured than Interview 2 and included more tasks and prompts. In Interview 2, there was only one main task with fewer prompts, and students explored varying numbers of predictions as they built their range of plausible values for the unknown parameter. Thus, I identified 18 reasoning clips in Interview 1 but the number of reasoning clips in Interview 2 varied. In APPENDIX E, I identified which prompt(s) were associated with each of the clips in Interview 1. In APPENDIX F, I identified the prompt(s) that were associated with Lorraine's clips, as an example, in Interview 2. As I analyzed participants' explanations and justifications line-by-line, I created memos for each participant in which I summarized their actions and responses to prompts, making initial hypotheses about the forms of reasoning they employed. In this first level of analysis, I stayed close to the raw data and identified small segments of the data (i.e., individual participant responses to prompts) that became the unit of analysis when applying Peirce's (1878) case-rule-result framework my second level of analysis.

In the second level of analysis, I applied Peirce's (1878) case-rule-result framework to each reasoning clip I identified in the first level of analysis by using Reid and Knipping's (2010) interpretation of Peirce's (1878) definitions of case, rule, and result. Following the analysis of Conner et al. (2014), I found it useful to first identify the rule within a reasoning clip. Determining its role in students' explanations informed my identification of the case and result. For each reasoning segment, I examined the structure of students' reasoning by identifying how they linked each case, rule, and result, thus inferring the form of reasoning they employed. During this level of analysis, I often reorganized the structure of case, rule, and result within a segment to compare multiple interpretations of the logical structure of the student's reasoning. In addition, it was useful to compare my analysis of students' reasoning across the same clips (in Interview 1) or similar clips (in Interview 2), checking for any inconsistencies in my interpretations. In this second level of analysis, I coded each reasoning clip with one or more of the three forms of inferential reasoning. This enabled me to identify patterns in students' reasoning, supporting me in the third level of analysis and my "development of explanatory constructs" (Simon, 2019, p. 120).

In the third level of analysis, I characterized students' reasoning more broadly and identified the role of deduction, induction, and abduction in students' understanding of sampling distributions. I used Transana to build reports that helped me visualize important patterns of reasoning across the participant pool. For example, examining reports from Interview 1 enabled me to see that students engaged in repeated iterations of inductive and deductive reasoning and that these iterations occurred after each time students drew samples. I then examined these patterns of reasoning more closely by generating reports of subsets of reasoning clips across all participants. For instance, I generated a summary report for five consecutive clips in Task A in

which the majority of students reasoned deductively. This report, separated by student, included all statements that students made during those five reasoning clips. I identified similarities and differences across students' deductive reasoning. Through this process, I identified emergent themes that provided additional insight into students' reasoning about and with sampling distributions beyond the three broader forms. This third level of analysis enabled me to make inferences about the role of the three forms of reasoning in students' understanding and characterize the more nuanced differences in students' reasoning.

Summary

The purpose of this chapter was to describe the data collection and analysis methods I used to answer my research questions. First, I justified the use of task-based clinical interviews, then provided the theoretical background and characteristics of the clinical interview methodology. Next, I described the participants I recruited for my study and then provided details about my data collection process, with a detailed description of the four tasks I used across both interviews. Lastly, I described my data analysis methods. These data collection and analysis methods were appropriate to examine the forms of inferential reasoning students employed when reasoning about and with sampling distributions.

CHAPTER 5

REASONING ABOUT SAMPLING DISTRIBUTIONS

The goal of Interview 1 was to examine how novice statistics students reasoned under uncertainty when given a population with a known population parameter. More specifically, I aimed to understand how students reasoned about the repeated sampling process, including how they reasoned about and estimated sampling variability. At the start of Interview 1, I gave students information about students at a large university with a known parameter of interest—North University reported on their website that 20% of their undergraduates are intended business majors. Additionally, I asked students to imagine asking a random sample of 100 undergraduates from North University if they intend to major in business. All of the subsequent prompts in all three tasks related to this sampling situation at North University. In Interview 1, students reasoned about sample outcomes as they completed three tasks, each with several prompts. Although all three tasks used the same North University context, they differed with respect to how students drew samples. Students did not draw any samples in Task A (see APPENDIX A), they drew physical samples from a box of beads in Task B (APPENDIX B), and they drew samples using a web-based simulation applet in Task C (APPENDIX C). These three tasks presented students with multiple opportunities to reason about sample outcomes from a population with a known parameter. These opportunities included providing an interval estimate for the outcome of a random sample of 100 undergraduates from North University, predicting the outcome of many hypothetical random samples of size 100, physically drawing and examining the outcomes of five and then 20 samples from a large box of beads with proportions

corresponding to North University, comparing predicted and observed sample outcomes, determining surprising outcomes, and using a web-based applet to simulate taking 500 random samples of size 100. Although most of the prompts differed across Tasks A, B, and C, some were repeated two or more times. I intentionally asked students to answer the same prompt multiple times to examine their reasoning across the different tasks. For example, I asked students to provide a range of values they would expect to get in a random sample of 100 undergraduates from North University at the end of Task A and then again at the end of Task C. This was intentional because I wanted to examine how students reasoned about sampling variability prior to drawing any samples (in Task A) and after drawing many physical and simulated samples (in Tasks B and C). Throughout Interview 1, students repeatedly reasoned about what sample outcomes they might expect and what sample outcomes they would find surprising. Thus, the results from Interview 1 help answer Research Question 1. Recall, in Research Question 1, I asked:

When given a population with a known parameter, what forms of reasoning do novice statistics students employ when determining what sample outcomes they expect and what sample outcomes they find surprising, and what do these forms of reasoning reveal about their reasoning *about* sampling distributions?

As I expected, students reasoned differently prior to drawing samples and after drawing samples. In general, students reasoned deductively in Task A (prior to drawing any samples) and inductively in Tasks B and C (after drawing samples). Within each of these broader forms of reasoning, I identified additional differences in patterns of reasoning across the student pool that revealed different understandings about sampling distributions. Thus, I organize the results from Interview 1 into two main sections: (1) Results prior to drawing any samples (Task A) and (2)

Results after drawing samples (Tasks B and C). In each section, I first give a description of the task(s) by listing specific prompts and activities in which students engaged. Next, I provide an overview of the major findings from the task(s), including describing differences in the broader forms of reasoning (i.e., deductive and inductive) across the student pool and identifying one or more additional categories of reasoning within these broader forms. Then I support the major findings by describing and comparing the reasoning of two or more representative students within each of the categories of reasoning within the broader forms. Throughout both sections, I identify what the broader forms of reasoning and the more nuanced differences in reasoning revealed across the student pool.

Prior to Drawing Samples: Results from Task A

Description of Task A

Recall at the beginning of Interview 1, I gave students information about students at a large university with a known parameter of interest—North University reported on their website that 20% of their undergraduates are intended business majors—and asked students to imagine asking a random sample of 100 undergraduates from North University if they intend to major in business. In Task A, students first provided an outcome for what they would expect to get in their random sample. Next, they compared the outcome of their hypothetical sample to a friend's hypothetical sample. After I presented a hypothetical sample outcome of 24%, I asked students if they would be surprised by this particular sample outcome and if they thought this particular sample indicated that the information posted on North University's website was incorrect. Lastly, they provided sample outcomes that would surprise them and gave a range of outcomes they would expect to get in a random sample of 100 undergraduates from North University.

Overview of Findings From Task A

In Task A, I expected students to reason based on informal facts or rules recalled from their recently completed introductory statistics course. For example, I expected that students would come to the study with some understanding that a sample is not the same as the population from which it was drawn and will not produce an outcome that is exactly equal to the population parameter. However, I did not expect students to use particular formulas to, say, calculate sampling variability or normal probabilities. The results from Task A aligned with these expectations in that all students had some intuition that sampling variability exists and none used formal calculations based on statistical formulas.

Regarding the three broader forms of reasoning (i.e., deductive, inductive, and abductive), students generally reasoned deductively throughout Task A with no evidence of reasoning inductively or abductively. This result is unsurprising because they had not yet drawn any physical or simulated samples. Thus, there was not a collection of observations and results from which they could generalize patterns. Although students generally reasoned deductively throughout Task A, I identified differences in reasoning across the student pool that revealed differences in their understanding of how far the sample statistic varies from the population parameter. At the level of the broader forms of reasoning, 8 out of the 11 students reasoned deductively to provide a reasonable range of values for what they would expect to get in a random sample of 100 undergraduates from North University; the other three did not reason deductively and provided an unreasonable range of values. A discussion of what is considered a reasonable range of values is forthcoming. For these eight students, deductive reasoning revealed that they came into this study with some understanding of how far a sample outcome varies from the population.

Of the eight students who reasoned deductively to provide a reasonable range of values, two showed evidence of a more sophisticated understanding of how far a sample varies from the population parameter because they chose their margin⁵ based on the sample size. The other six did not show evidence of this understanding because they chose their margin based on a personal feeling or a fact they recalled from a previous course. The three students who provided an unreasonable range of values gave ranges that were too wide (e.g., 0% to 80%), suggesting an overreliance on sampling variability, what Rubin et al. (1991) described as the inclination that a random sample tells you nothing about the population; the sample statistic could be anything. For these three students, including extremely low and high values, such as 0% and 80%, shows that they do not yet show evidence of understanding how far the sample statistic varies from the population parameter.

In summary, the results of Task A revealed three categories of reasoning across the student pool: (1) students who reasoned deductively to provide a reasonable range of values and chose their margin based on the sample size, (2) students who reasoned deductively to provide a reasonable range of values and chose their margin based on a personal feeling, intuition, or previous knowledge, and (3) students who did not reason deductively and provided an unreasonable range of values. Table 4 identifies the category of reasoning for each of the 11 students.

Using Deductive Reasoning to Provide a Reasonable Range of Values

Eight students used a particular margin to deduce a range of sample outcomes they would expect for a random sample of 100 undergraduates from North University. These students provided an interval estimate that was symmetric around 20% with a margin between 5% and

⁵ Although some students called this value “margin of error,” I refrain from using this phrase because it is used to represent the range of uncertainty or variability around a statistic, not a parameter, as in this case.

Table 4*Categories of Reasoning in Task A*

Form of Reasoning	Activity	Student										
		Tami	Julie	Jess	Becky	Lorraine	Tyra	Waverly	Eric	Lyla	Mindy	Corrina
Reasoned deductively to provide a reasonable range of sample outcomes centered at 20%	Chose margin based on a personal feeling, intuition, or previous knowledge	X		X			X		X	X	X	
	Chose margin based on sample size		X			X						
Did not reason deductively	Provided an unreasonably wide range of sample outcomes not centered at 20%				X			X				X

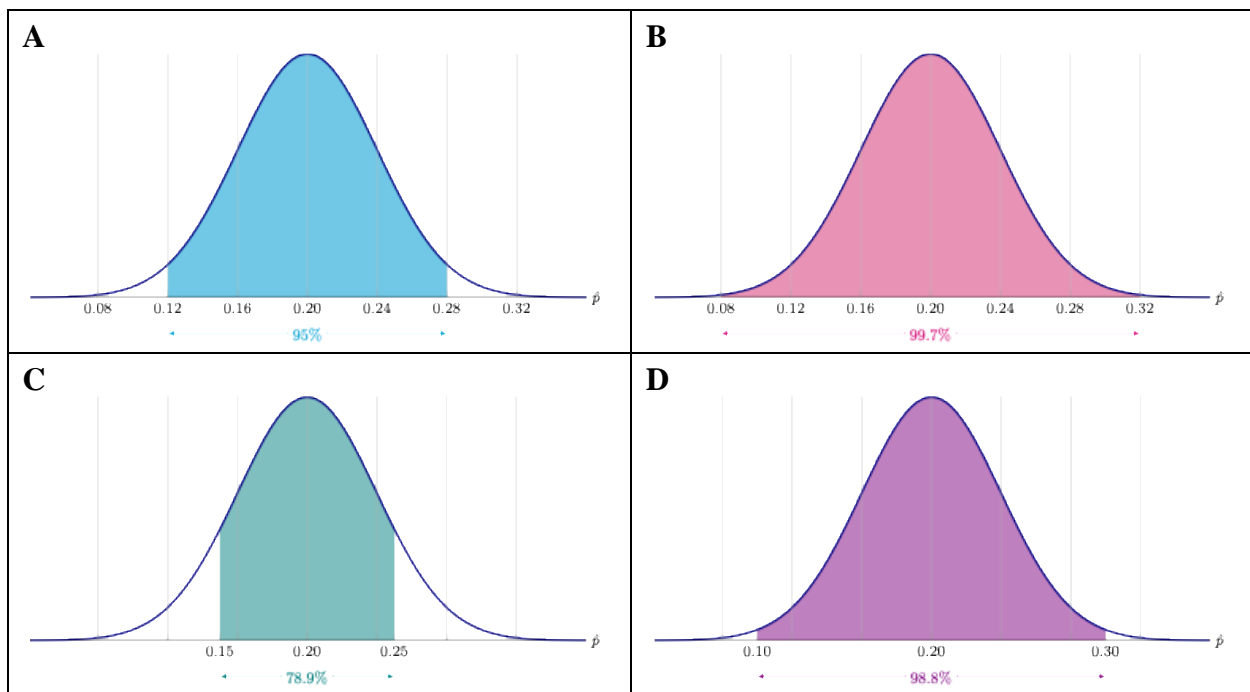
10%. Although not all eight students used the same margin or the same justification for their choice of margin, each of them applied a rule (their chosen margin) to a case (a random sample of 100 undergraduates drawn from North University) to deduce a result (a range of sample outcomes). These results show that these students understand that the sample outcomes should be centered at the population parameter and that the sample outcomes vary from the population parameter. Furthermore, they did not provide an unreasonably wide range of outcomes or a range that extended to 0% or 100%. By giving a margin of between 5% and 10%, these students gave a reasonable estimate for how far they expected a sample outcome to vary from 20%. This indicates that they understand that the sample outcomes would not vary so much from the population parameter such that any sample outcome is possible. Before illustrating this form of reasoning with excerpts from students, I discuss what is considered a reasonable range of values in the context of North University. It is important to note that the following discussion was not part of the interview protocol, nor did I expect students to assess the normality of the sampling

distribution, calculate the standard deviation of the sample statistic, or use normal probabilities to construct their range of values.

What is Considered Reasonable? What constitutes a *reasonable* range of values is certainly subjective, but considering the normality of the sampling distribution for this context is helpful in understanding how far the sample statistic varies from the population parameter. In the context of North University, the sampling distribution of the sample statistic—the proportion or percent of intended business majors in a random sample of 100 undergraduates—is approximately normal with a mean of 0.2 (or 20%) and a standard deviation of 0.04 (or 4%). Using normal calculations, approximately 95% of all possible random samples of 100 undergraduates from North University should produce a proportion of intended business majors between 0.12 and 0.28, the values that are two standard deviations from the mean (see Figure 8a). Approximately 99.7% of the samples should produce outcomes between 0.08 and 0.32, the values that are three standard deviations from the mean (see Figure 8b). A reasonable estimate is a range of outcomes that encompasses *most* of the data. Again, this is a subjective statement. However, these normal calculations can help assess ranges that are more reasonable than others. For example, a range of 0.16 to 0.24 with a margin of 0.04 encompasses about 68% of the sample outcomes and could be considered less reasonable than 0.12 to 0.28 with a margin of 0.08 which encompasses about 95% of the sample outcomes. This is not to say a wider range is always more reasonable. For example, a range of 0 to 1 (or 0% to 100%) is wider than 0.12 to 0.28 but is unreasonable because the probability of observing a sample outcome close to 0 or close to 100 is essentially 0, or nearly impossible. A range with a margin of 5% encompasses approximately 78.9% of the sample outcomes (see Figure 8c), whereas a range with a margin of 10% encompasses approximately 98.8% of the sample outcomes (see Figure 8d).

Figure 8

Theoretical Sampling Distribution for the North University Context



Thus, the eight students who used a margin between 5% and 10% to deduce their range of sample outcomes provided ranges that were reasonable. Although all eight of these students deduced a reasonable range of sample outcomes, their justification for their choice of margin differed; six of the eight students chose their margin based on a personal feeling, intuition, or previous knowledge, whereas two chose their margin based on sample size. These two different justifications for their choice of margin provided valuable insight into students' understanding of how far a sample statistic varies from the population parameter. I discuss these two categories of reasoning in the next two sections.

Justifying the Margin Based on a Personal Feeling or Previous Knowledge. Of the eight students who deduced a reasonable range of sample outcomes, six justified their margin with a personal feeling or previous knowledge. For example, Tami chose a 5% margin because “5 is a good number.” Lyla explained her choice of a 10% margin and said, “To keep my 10%

inklings...between the 10 percents make sense in my head.” Eric relied on his knowledge from previous courses to justify why he chose a margin of 5%. He drew on his previous experiences in both chemistry and statistics:

One of the rules that you can kind of use sometimes is as long as your margin of error is within 5%, then usually you didn't do anything wrong. I know when I used to do...chemistry, as long as you were within 5%...if you're assigned a study and they have this number as the assumed, like when you cook this compound it should weigh about this. As long as you're within 5% of that then your margin of error isn't wildly off. Five percent's 0.05. With alphas, too, in statistics, so when you do P values and stuff like that, 0.05 is a good base.

Regardless of their specific margin, Tami, Lyla, Eric, and the three others who reasoned in this way applied their margin (rule) to a random sample of 100 undergraduates drawn from North University (case) to deduce a range of sample outcomes (result). Although these six students deduced a reasonable margin, their justification for their choice of margin was somewhat arbitrary and disconnected from the specific sampling situation at North University. In other words, these students did not connect their choice of margin to any of the constraints (e.g., sample size) in the North University context.

Justifying the Margin Based on the Sample Size. Of the eight students who deduced a reasonable range of sample outcomes, only two justified their margin based on the sample size. Like the other six students who provided a reasonable range of values, both Julie and Lorraine deduced their range of values from a rule. However, their deduction differed from the others' because their rule related sampling variability to sample size: Smaller samples produce sample outcomes that are more variable than larger samples (rule). They applied their rule to a sample

size of 100 (case) to deduce their range (result). For example, Julie chose 10% instead of 5% based on the sample size of 100, which she considered to be small. She explained, “The sample size is so small. If the sample size was much larger, I would probably have less variation. It would be within 5%.” This is evidence that Julie understands that larger samples produce outcomes with less variability from the parameter than smaller samples. Lorraine showed evidence of this understanding, too, when she repeatedly mentioned the sample size throughout Task A and used the sample size in her justification for why she chose a margin of 6-7%. Prior to giving her range of outcomes, Lorraine said that a sample outcome of 24% would not indicate that the information posted on North University’s website was incorrect, and specifically talked about the sample size in her explanation:

It doesn't because it's a, I think large enough sample to get reasonably similar results to however many students are business majors in the entire university. But it's probably not large enough to be so specific that if 24% said that they were business majors, it would just be totally like the school's percentage would be wrong.

Lorraine understands that a sample size of 100 is large enough to be representative of the population, but not so large that the sampling variability is so small that a 4% difference between the statistic and parameter is significant. When she explained why she chose a margin of 6-7%, she mentioned the sample size again. She said, “It's a large enough sample, but it's not a large sample, and so I think there could be much more variability in the mean percentage of business majors.” Lorraine’s thinking about the effect of sample size is quite sophisticated. Not only does she understand that larger samples are more representative of the population than smaller samples, but she also understands that larger samples produce more precise (less variable) estimates for the population parameter than smaller samples.

Providing an Unreasonable Range of Values

Unlike the previously discussed eight students, three students did not show evidence of reasoning deductively from a rule, nor did they provide a reasonable range of values. Instead, these students provided ranges that were too wide and not symmetric around 20%. Becky gave an interval of 0% to 80%, Waverly gave an interval of 10-15% to 50%, and Corrina gave an interval of 5% to 40%. These results indicate that these students did not understand that the sample outcomes would vary above and below 20% by same amount. In other words, they expected the sample outcomes to vary farther above 20% than below 20%. Furthermore, the outcomes that they deemed surprising were very far above and below 20%. For example, Becky said she would not be surprised by a sample outcome of 50% but would be surprised by sample outcomes greater than equal to 90% and then gave a range of 0% to 80% for what she would expect to get in a random sample of 100 undergraduates from North University. She explained,

Just from a random sample, not exactly sure how big of a population we're talking about.

So, we may not find, hypothetically, we may not find any of those business majors out of these 100 people. So lowest range would be 0 to 1, that's going to go to 0 as the lowest.

And highest, I guess 80 ish maybe? Or, yeah, I'll just go with 80. Kind of like what I said here, I would be really surprised to see at least more than 90, but I feel like this is a pretty realistic maximum amount for a single sample if it was completely random.

Becky's explanation indicates that she thinks that any sample outcome, from 0% to 100%, is possible and justified her reasoning based on the sample being random. At the beginning of A, all three students said they would expect to get around 20 (or 20%) business majors if they were to ask 100 randomly selected undergraduates from North University. In addition, all three students said they would expect a friend's hypothetical sample to also report close to 20%. These

responses paired with their unreasonable ranges suggest that these three students had difficulty balancing two important ideas about sampling: sample representativeness, the idea that a sample will often have similar characteristics as the population from which it was drawn, and sample variability, the idea that multiple samples from the same population are not all the same as each other, nor are they the same as the population from which they were drawn (Rubin et al., 1991).

After Drawing Samples: Results from Tasks B and C

Description of Tasks B and C

Throughout Tasks B and C, students repeatedly drew random samples, both physically from a box of beads and virtually using a web-based simulation tool. In Task B, students first predicted the outcome of five hypothetical random samples of 100 undergraduates from North University. Next, they physically drew five random samples of 100 beads from a large box of beads with proportions corresponding to North University's undergraduate population—20% of the beads were white and 80% were green (see Figure 9). I provided each student with parallel

Figure 9

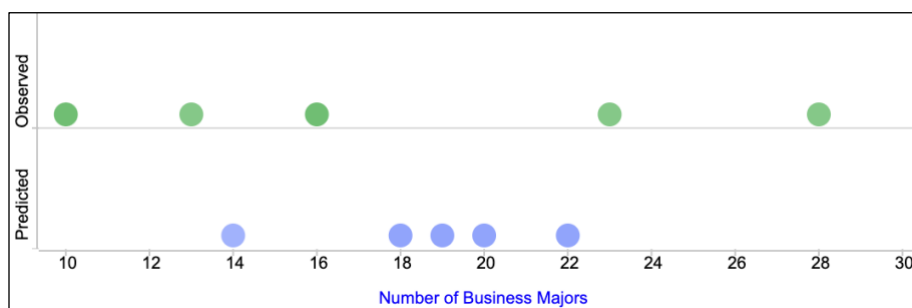
Box of Beads From Which Students Drew Samples in Task B



dot plots displaying their predicted outcomes and observed outcomes on the same scale (see Figure 10) and students described similarities and differences they noticed. They repeated this process (predict outcomes, draw samples, compare predicted and observed outcomes) with 20 random samples of 100. Lastly, students decided if they would be surprised by two particular sample outcomes—24% and 38%—and identified any other sample outcomes that would be surprising.

Figure 10

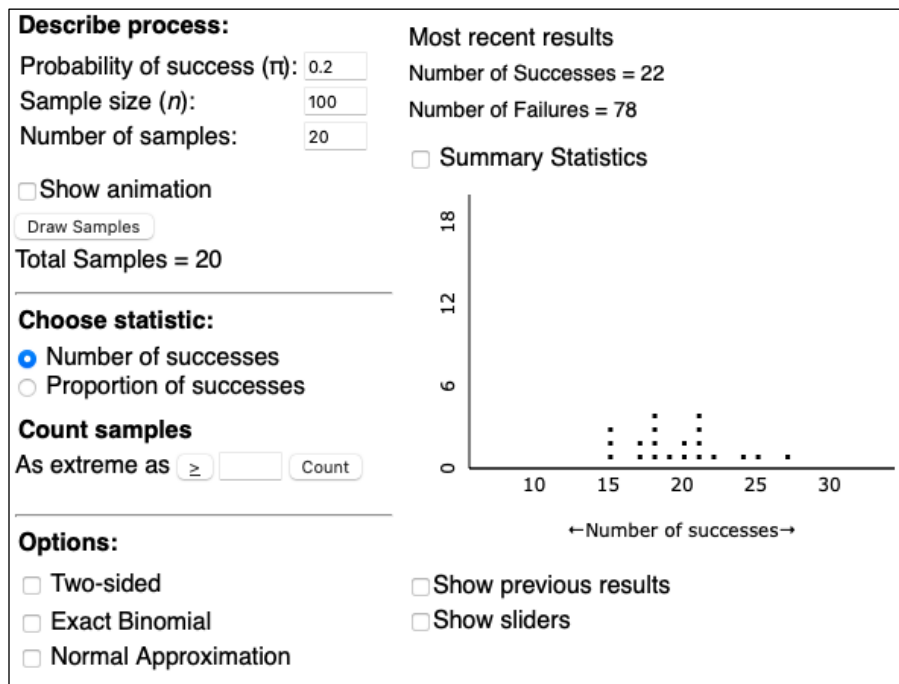
Parallel Dot Plots Displaying Predicted and Observed Outcomes



In Task C, I presented students with a web-based applet that they used to model the repeated sampling process in the context of North University. Figure 11 shows a screenshot of the applet displaying the results of 20 random samples of 100 from a population whose parameter is 20%. After a short introduction to the applet, I asked students to explore how it worked by recreating the repeated sampling process that they performed in Task B with the box of beads. Students used the applet to draw five random samples of 100, then 20 random samples of 100, then described what they noticed about the distribution of sample outcomes. Students repeated this process (draw samples, describe what you notice) with 500 random samples of 100. Like in Task A, students then gave a range of sample outcomes they would expect to get in a random sample of 100 undergraduates from North University. In this task, I provided students with their initial range from Task A and asked them to explain what changed in their thinking.

Figure 11

Screenshot of Web-Based Applet That Models the North University Context



Lastly, students identified what outcomes surprised them and drew a picture that represented the repeated sampling process that the applet was modeling for the North University context.

Overview of Findings From Tasks B and C

In Tasks B and C, I expected students to use the results of their samples to continually adjust and refine what outcomes they expected and found surprising in a random sample of 100 undergraduates from North University. For example, I expected that students would adjust their predictions for 20 random samples based on the results of the five random samples they drew from the box of beads. Similarly, I expected that students would adjust what outcomes they found surprising based on the results of the 20 random samples they drew from the box of beads. I also expected that students would refine their range of values from Task A based on the results of the 500 samples they drew using the web-based applet. The results from Tasks B and C partially aligned with these expectations in that all but one student did use the results of their

samples to continually adjust and refine what outcomes they expected and found surprising. However, there were several differences in students' reasoning throughout Tasks B and C that provide insight into their understanding about sampling distributions.

Regarding the three broader forms of reasoning (i.e., deductive, inductive, and abductive), 10 students used a combination of inductive and deductive reasoning to determine what sample outcomes they might expect and what sample outcomes they would find surprising. They repeatedly generalized one or more patterns from the observed sample outcomes to a larger group, then used those patterns and generalizations to inform their thinking about a later collection of sample outcomes. Like in Task A, these students attended to sampling variability throughout Tasks B and C. However, their reasoning about sampling variability was more sophisticated in Tasks B and C in that they moved beyond talking about the *existence* of sampling variability to trying to *estimate* sampling variability. Their repeated iterations of inductive and deductive reasoning helped them to continually refine their estimate for sampling variability, supporting their understanding of how far a sample outcome varies from the population parameter.

Surprisingly, one student—Tyra—never used the results of her previously drawn samples to adjust or refine what sample outcomes she would expect or find surprising. At the end of Task A, Tyra used a reasonable margin of 5% to deduce her range of 15% to 25% and based her choice of margin on a personal feeling. Throughout Tasks B and C, there was no visual or verbal evidence that Tyra used the results of any of the samples she had drawn to inform her thinking about what she would expect or find surprising in future outcomes. Instead, she reasoned similarly to how she reasoned in Task A, deducing what she would expect or find surprising based on a personal feeling for what she thought was “close to” or “far from” 20%. Furthermore,

Tyra gave the same range of values at the end of Task C that she did at the beginning of the interview, choosing the same margin of 5% based on a personal feeling. One possible explanation for this is that Tyra did not view the box of beads or the web-based applet as models of the North University context. Another explanation is that the results of Tyra's samples did inform her thinking, but her words and actions did not convey that she used the results of her samples in her reasoning. At this point there is no more evidence for one explanation than the other, but I will discuss Tyra's reasoning in more detail in a later section.

In my analysis of the other 10 students' reasoning in Tasks B and C, I identified four components that were important in understanding how they reasoned about the repeated sampling process, including how they reasoned about and estimated sampling variability:

- (1) Students compared the distributions of predicted and observed sample outcomes using characteristics like shape, center, or spread;
- (2) Students assessed and compared the probabilities of one or more sample outcomes by using words or phrases such as "likelihood," "more probable," "very unlikely," and "improbable";
- (3) Students provided an updated range at the end of Task C that was both reasonable and centered at 20%;
- (4) Students justified their range based on the 500 samples they had drawn using the web-based applet or based on all samples they had drawn, both from the box of beads and using the applet.

Components 1 and 2 provide evidence of thinking about the collection of sample outcomes as one entity—or an aggregate—rather than focusing on individual outcomes. Component 3 indicates an understanding of how far the sample outcomes should vary from 20%. Component 4

suggests an understanding that the collection of the 500 sample outcomes represents the possible outcomes for one random sample of 100 undergraduates from North University. Taken together, these components contribute to a more sophisticated understanding of sampling distribution than each component separately. For example, students who included only Component 1 or Component 2 in their reasoning showed some evidence of an aggregate view of data, but without providing a reasonable range of sample outcomes based on the results of many (500 or more) random samples, these students did not show an understanding that the repeated sampling process produces a collection of possible sample outcomes and that one sample outcome is simply one such case. Table 5 displays the categories of reasoning I identified for Tasks B and C. Within each form of reasoning, I identified which components each student included in their reasoning. Although Tyra showed evidence of Component 3 in her reasoning in Tasks B and C, because she never reasoned inductively, I discuss her reasoning separately from the 10 students who did reason inductively. Thus, my subsequent categorizations and descriptions are for those 10 students.

From Table 5, seven students showed evidence of at least three components: Components 1 or 2 (or both), Component 3, and Component 4. These components suggest a critical understanding about sampling distribution—that a sample is simply one case out of a collection of similar cases. Although the other three students reasoned both inductively and deductively to update what outcomes they would expect and find surprising, either they did not explicitly consider the likelihood of particular sample outcomes in their reasoning, or they focused on individual outcomes when comparing the distributions of their predicted and observed sample outcomes. Furthermore, at the end of Task C, they gave an updated range of values that was not

Table 5*Categories of Reasoning in Tasks B and C*

Form of Reasoning	Component	Student										
		Tami	Julie	Jess	Becky	Lorraine	Tyra	Waverly	Eric	Lyla	Mindy	Corrina
Inductive	Compared shape, center, or spread in distributions of predicted and observed sample outcomes	X		X	X	X		X		X	X	
	Compared individual outcomes in distributions of predicted and observed sample outcomes		X									X
	Assessed and compared the probabilities of one or more sample outcomes	X		X		X			X	X	X	
	Provided an updated range of outcomes that was reasonable and centered at 20%	X		X		X		X	X	X	X	
	Justified updated range based on the results of the 500 simulated samples or all samples, both physical and simulated	X		X		X		X	X	X	X	
Deductive	Compared individual outcomes in distributions of predicted and observed sample outcomes						X					
	Provided an updated range of outcomes that was reasonable and centered at 20%						X					

centered at 20%, but more importantly, they focused on individual sample outcomes in the smaller collections of samples—five and 20—that they drew from the box of beads when they justified their updated range, rather than the 500 sample outcomes from the web-based applet. This was surprising because I expected students to justify their range based on the 500 samples they had drawn using the web-based applet or based on all samples they had drawn, both physical and simulated. One explanation for this result is that these students thought that a smaller collection of sample outcomes, like five or 20, was more helpful in estimating a range of

outcomes for one random sample rather than 500 samples. This would indicate that these students did not view the collection of 500 sample outcomes as the best approximation for the theoretical sampling distribution from which they could estimate probabilities. Another explanation is that these students viewed the simulation tools differently. For example, it is possible that these students had more confidence in the samples they drew from the box of beads because they had some control over the repeated sampling process. In contrast, they simply clicked buttons on the web-based applet and were not exposed to the probability mechanisms built into the applet or the coding involved in making the applet function, which may have caused skepticism in the trustworthiness of the applet.

The results of Tasks B and C revealed three categories of reasoning across the student pool: (1) students who reasoned both inductively and deductively and showed evidence of at least three of the four components identified previously, (2) students who reasoned both inductively and deductively and showed evidence of at most one of the four components, and (3) one student who reasoned deductively and did not consider the results of any of the samples drawn from the box of beads or using the applet.

Using Inductive and Deductive Reasoning to Continually Adjust and Refine Expected and Surprising Outcomes

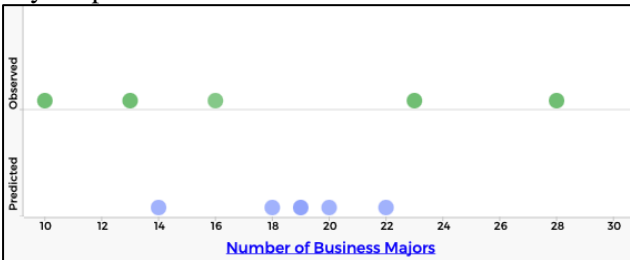
Ten students used a combination of inductive and deductive reasoning multiple times to inform their thinking about later sample outcomes. In Task B, these students used one or more patterns they noticed in the five observed sample outcomes from the box of beads to predict what they would expect to get in 20 hypothetical random samples. They repeated this pattern of reasoning after drawing 20 random samples from the box of beads, using the patterns they observed in previous sample outcomes to adjust what they expected and found surprising. In

task C, these students used one or more patterns they noticed in previously drawn samples (physical or simulated) to update their range of values for what they would expect to get in a random sample of 100 undergraduates from North University. These students observed one or more patterns (result) in several observed sample outcomes (cases) and generalized that pattern to a larger group (rule). Then they applied their generalization (rule) to a new collection of sample outcomes (cases) to determine what they expected or found surprising (result).

Lorraine's reasoning is typical of the group; Table 6 shows one example of her inductive and deductive reasoning in Task B when she used the results of her five observed samples to inform her thinking about what she would expect to get in 20 sample outcomes. In the table, I list four prompts, Lorraine's response to each prompt, and an explanation of my analysis of her reasoning. Prior to this reasoning excerpt, Lorraine had just predicted what she would expect to get in five random samples. She then drew five random samples of 100 beads from the large box of beads. The excerpt in Table 6 begins after I gave Lorraine parallel dot plots displaying her predicted and observed outcomes on the same scale. The first row of the table shows a recreation of the parallel dot plots, for better visibility. Lorraine compared her predictions with the observed sample outcomes and noticed "a bit more spread" than what she initially predicted at the end of Task A. Lorraine identified the minimum and maximum of the observed outcomes at 10 and 28, respectively. She observed a pattern in her five observed sample outcomes—the minimum is 10 less than 20, the expected value, and the maximum is eight more than 20—and generalized this pattern to adjust what she thought was a reasonable amount of variability from 20. In short, Lorraine expected that what she noticed in five sample outcomes would also occur in 20 sample outcomes. She used the results of her five random samples to adjust her original estimate from 6-7% to 8-10% for how far the sample outcomes would vary from the population parameter. Then

Table 6

Lorraine's Inductive and Deductive Reasoning in Task B

Prompt	Response	Analysis																																												
<p>The parallel dot plots display the actual sample outcomes and your predicted sample outcomes. How do the actual sample outcomes compare to your predictions?</p> 	<p>“It seems like the actual observed sample outcomes are, there's a bit more spread in the results. So, the lowest was 10 and the largest was 28 and I was giving a range of around six or seven to be reasonable, but this seems like the range is eight to 10 around that from the initial 20% in terms of what would a reasonable percentage of white beads be to get out of the box. But it seems like there's just a bit of a wider spread.”</p>	<p>Lorraine observed that the five random samples (cases) produced sample outcomes that ranged from 10 to 28 (results). She used her results from five random samples to generalize what she expected to be a reasonable amount of variability from 20%. She induced that a reasonable range of percentages of white beads would be eight to 10 around the initial 20% (rule).</p>																																												
<p>Predict the outcome of 20 different random samples of 100 undergraduates from the population of all undergraduates at North University.</p>	<table><tr><td>Sample</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td></tr><tr><td>Prediction</td><td>17</td><td>12</td><td>25</td><td>19</td><td>13</td><td>22</td><td>23</td><td>18</td><td>20</td><td>14</td></tr><tr><td>Sample</td><td>11</td><td>12</td><td>13</td><td>14</td><td>15</td><td>16</td><td>17</td><td>18</td><td>19</td><td>20</td></tr><tr><td>Prediction</td><td>16</td><td>21</td><td>29</td><td>15</td><td>22</td><td>13</td><td>19</td><td>26</td><td>27</td><td>20</td></tr></table>	Sample	1	2	3	4	5	6	7	8	9	10	Prediction	17	12	25	19	13	22	23	18	20	14	Sample	11	12	13	14	15	16	17	18	19	20	Prediction	16	21	29	15	22	13	19	26	27	20	<p>Lorraine used this updated range to predict what she would expect to get in 20 random samples. She applied her rule to 20 hypothetical random samples (cases) to deduce the sample outcomes she would expect to get (results).</p>
Sample	1	2	3	4	5	6	7	8	9	10																																				
Prediction	17	12	25	19	13	22	23	18	20	14																																				
Sample	11	12	13	14	15	16	17	18	19	20																																				
Prediction	16	21	29	15	22	13	19	26	27	20																																				
<p>Explain how you made your predictions.</p>	<p>“I predicted quite a few more numbers that are a bit further away from 20 kind of in the 5, 6, 7 away range, just because that seems more, based on the last results that we got, that seems more likely than I initially thought. So, there's a bit more, the results are a bit more varied.”</p>																																													
<p>What, if anything, changed in your thinking from your first set of predictions (for the five random samples of 100) to this second set of predictions (for these 20 random samples of 100)?</p>	<p>“Just because I know that maybe it's more likely that there'll be numbers that are further away from 20 like 10 or 28 from the last one. Yeah, same thing. I just spread out the results a bit more. They're less concentrated around 20.”</p>																																													

she used her updated estimate to inform what she would expect to get in 20 random samples. After she predicted the outcome of 20 random samples, she drew 20 random samples from the box of beads, compared her predictions to the actual outcomes, then engaged in another iteration of inductive and deductive reasoning to identify what sample outcomes would be surprising. To visualize how Lorraine updated her estimate for sampling variability, I present the parallel dot plots of her predicted and observed outcomes for five and 20 random samples in Figure 12.

Figure 12

Lorraine's Predicted and Observed Sample Outcomes



From the plots, the range (distance from the minimum value to the maximum value) of Lorraine's 20 predictions is roughly equal to that of her five observed sample outcomes (17 compared to 18). More importantly, the amount that Lorraine's 20 predictions vary from the expected value of 20 is also roughly equal to that of her five observed sample outcomes (8-10). Recall, Lorraine and nine others engaged in repeated iterations of inductive and deductive reasoning like the one I described in Table 6. By engaging in these forms of reasoning, Lorraine and the other nine students continually refined what sample outcomes they expected and found surprising. In doing so, they developed a better understanding of how far the sample statistic varies from the population parameter of 20%. This was especially evident when Becky, Waverly, and Corrina—the three students who gave unreasonably wide ranges in Task A—provided

updated ranges that were much more narrow and more reasonable at the end of Task C. For example, Becky initially gave a range of 0% to 80% at the end of Task A and adjusted her range to 5% to 32% by the end of Task C. Even though all 10 of these students developed a better understanding of sampling variability by the end of Task C, the differences in their reasoning within these broader forms provide additional insight into their ways of thinking about and understanding sampling distributions.

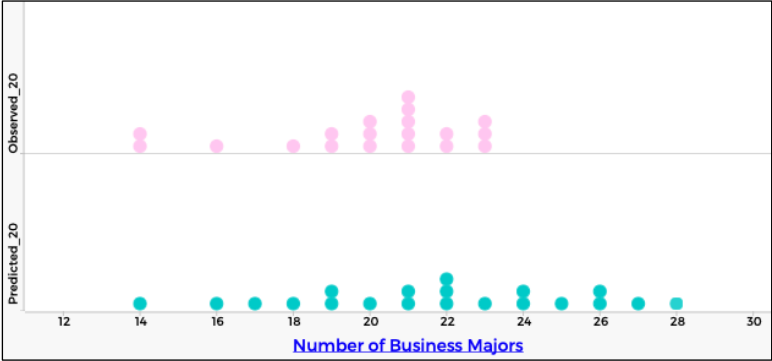
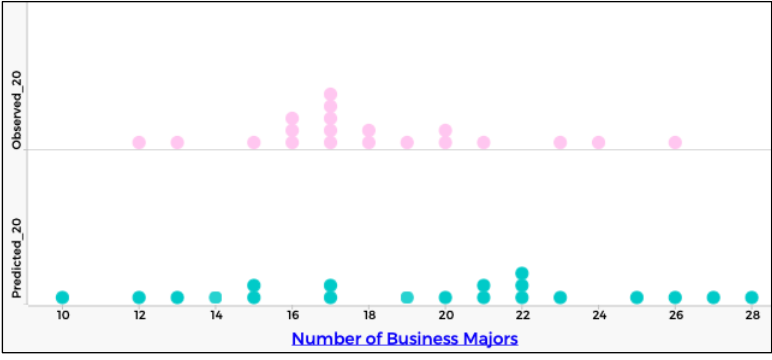
Providing a Reasonable Updated Range of Values. Of the 10 students who repeatedly reasoned inductively and deductively to adjust their estimates for sampling variability, seven included at least three of the four components in their reasoning throughout Tasks B and C, indicating they had a distributional perspective of sample (Kahneman & Tversky, 1982), one in which they viewed one sample as one case out of many possible cases, for which probabilities can be estimated. In other words, they showed evidence of understanding that the outcome of one random sample of 100 undergraduates drawn from North University could be any one of the five, 20, or 500 samples they drew from the box of beads or using the web-based applet.

These seven students included at least one of Components 1 and 2 in their reasoning, indicating they viewed collections of sample outcomes as one entity, or an aggregate. Component 1 shows evidence of viewing data as an aggregate because aspects like shape, center, and spread are characteristics that identify and summarize patterns in the entire data set (Rubin, 2020). The ways in which both Jess and Waverly compared the distributions of their predicted and observed sample outcomes for 20 random samples is typical of these seven who used Component 1 in their reasoning. After predicting the outcome of 20 random samples then drawing 20 random samples from the box of beads, I provided students with parallel dot plots displaying the distributions of their predicted and observed outcomes on the same scale.

Error! Reference source not found. shows a recreation of the parallel dot plots I provided to Jess and Waverly and their comparison of their predicted and observed outcomes. Jess compared the spreads of the two distributions; she noticed that her predicted outcomes were “a little bit more spread out” with “a wider variety of numbers” than her observed outcomes that were “a little more compact” and “more together.” She also compared the shapes of the two distributions; she identified the distribution of the observed outcomes as somewhat “bell-shaped” in contrast to the shape of the distribution of her predicted outcomes, which she identified as “skewed.” Jess compared two aspects—shape and spread—of her predicted and observed outcomes. Like Jess, Waverly also compared the spreads of her predicted and observed outcomes when she noticed the observed outcomes were “not as spread out” and “a little bit more narrow.” However, Waverly described the spread in terms of the variability from the center when she said, “The lower range and the higher range are closer to 20 than I expected.” Though not explicitly, she implied that she compared the centers by comparing the variability from 20. Both Jess and Waverly compared two aspects of the distributions; Jess compared the spreads and shapes, whereas Waverly compared the variability from the center indicating she compared the spreads and centers. Jess and Waverly compared aspects of the distributions that summarize the entire collection of outcomes, indicating they were able to view the collection of outcomes as a whole. According to Ben-Zvi (2004), a *local understanding* of data focuses on individual observations in a data set, whereas a *global understanding* involves perceiving the data as an entity and being able to identify and recognize patterns. Students like Jess and Waverly who compared their predicted and observed outcomes based on characteristics like shape, center, or spread demonstrated evidence of a global understanding of data and viewing data as an aggregate by identifying and recognizing patterns in their predicted and observed outcomes.

Table 7

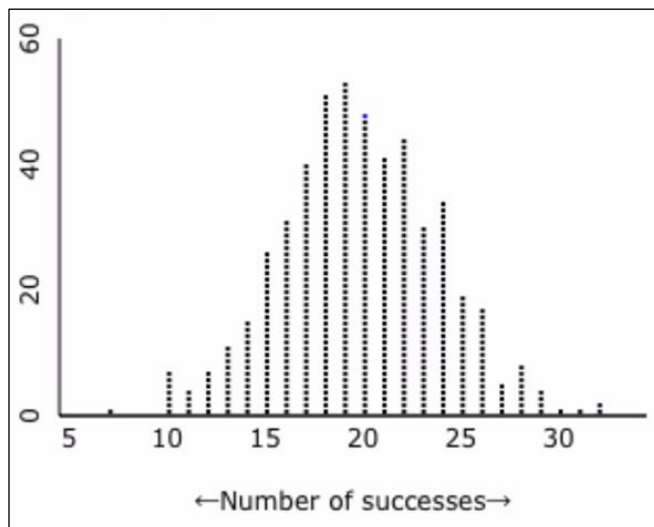
Comparing Predicted and Observed Outcomes

Participant	Prompt	Response
Jess	<p>The parallel dot plots display the actual sample outcomes and your predicted sample outcomes. How do the actual sample outcomes compare to your predictions?</p> 	<p>“My predicted outcomes are a little bit more spread out versus my observed a little more compact, like that one outlier. Kind of more like, not really bell-shaped, but a little bit small bell, versus my observed, it's skewed...I feel like my observed is more together versus my predicted had a wider variety of numbers.”</p>
Waverly	<p>The parallel dot plots display the actual sample outcomes and your predicted sample outcomes. How do the actual sample outcomes compare to your predictions?</p> 	<p>“They're not as spread out. The lower range and the higher range are closer to 20 than I expected them to be. So, it doesn't have as much variability as I was expecting. Because the observed, I think the lowest was 12, and then the highest was 26 and I had 10 and 28. So it's just a little bit more narrow than I expected it to be.”</p>

Component 2 also provides evidence of viewing data as an aggregate because assessing and comparing the probability of one or more sample outcomes involves thinking about relative frequency, that is, comparing the frequency of one outcome to the collection of all outcomes. For example, Tami assessed the probability of an individual sample outcome and a range of sample outcomes. In Task B, she determined she would be surprised if a random sample of 100 undergraduates from North University reported 38 intended business majors by assessing the probability of observing a sample outcome of 38. While looking at the dot plot displaying the results of the 20 random samples she drew from the box of beads, she noticed that the highest sample outcome was 23 and noted that “38 is a whole 15 numbers away, so very, very unlikely.” Tami’s reasoning about the probability of getting a sample outcome of 38 shows that she assessed the relative frequency of 38. In other words, Tami compared the frequency of 38 to the collection of all 20 sample outcomes, evidence that she viewed the collection of outcomes as a whole. At the end of Task C, Tami assessed the probability of a range of sample outcomes when she said she would expect to get between 16 and 24 intended business majors in a random sample of 100 undergraduates from North University. While looking at the dot plot displaying the results of the 500 random samples she drew using the web-based applet, she gave her range of 16 to 24 and noted that “in between that range is where the highest peak of the bell curve is.” When I prompted Tami to explain further, she said, “That just means that it’s very likely to get a number from there.” Like Tami, Eric also assessed the probability of an individual sample outcome. In Task C, after he drew 500 random samples using the applet, Eric noticed what he considered to be an outlier. Figure 13 shows the distribution of Eric’s 500 sample outcomes; he identified the seven as an outlier.

Figure 13

Results of Eric's 500 Random Samples



I asked Eric how he could explain getting an outcome that low if he assumed he took a truly random sample. After a 15-second pause he responded,

I mean I guess you can attribute it back to, out of 500, that's one...So, when I see that I'm like, oh, the chances are really small. And obviously the higher that number gets, like if this was a thousand [samples] and there was one (points to the outcome of seven), then I would expect that a lot less. But I mean obviously it's still a possibility, but the likelihood is obviously a lot less likely than getting like 20 or 19.

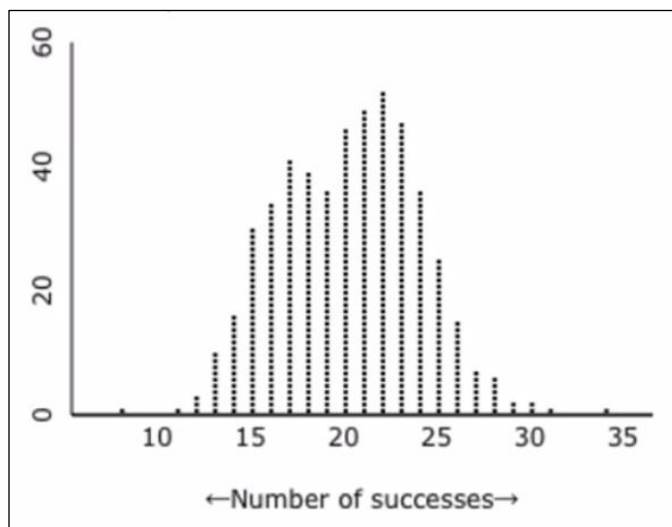
To assess the likelihood of the outcome he identified as an outlier, Eric explicitly identified the relative frequency of the outcome (1 out of 500) and explained that the outcome would be even less likely if the relative frequency were 1 out of 1000. Furthermore, he compared the likelihood of getting a sample outcome of seven to that of getting an outcome of 20 or 19. Eric's assessment and comparison of one or more probabilities showed that he related the frequency of one or more outcomes to the collection of all 500 samples outcomes. Students like Tami and Eric who attended to the probability of one or more sample outcomes demonstrated that they reasoned

about relative frequencies, which involves comparing individual outcomes to the distribution of all outcomes, supporting a more global view of the data (Bakker & Gravemeijer, 2004).

Recall, seven of the 10 students who reasoned both inductively and deductively in Tasks B and C included both Components 3 and 4 in their reasoning. Although providing an updated range of values that is both reasonable and centered at 20% (Component 3) indicates that these students understand how far the sample outcomes should vary from 20%, their justification for their range provides important insight into how they reasoned about sampling distributions. Rather than simply reporting students' updated ranges, I will discuss students' updated ranges and their justifications together. Recall, all seven justified their updated range of values based on the results of the 500 samples they drew using the applet or all samples they drew, both physical and simulated (Component 4). For example, Mindy updated her initial range of 10 to 30 at the end of Task A to 15 to 25 at the end of Task C and justified her range based on the results of the 500 samples she had drawn using the web-based applet. Figure 14 shows the distribution of Mindy's 500 sample outcomes.

Figure 14

Results of Mindy's 500 Random Samples



The excerpt below begins with Mindy giving a range of outcomes she would expect to get in any one single random sample of 100 undergraduates from North University.

Mindy: (Looks at dot plot displaying the results of her 500 random samples.) I think you would expect to see, if you just get one random sample...I think with one it could be like...I think anything from, what, 15 to 25 is more common than anything else. But we did see some farther away, but not as often clearly.

Claire: So, if you were to go take one single random sample of 100 undergrads from North University, you think it might be between 15 and 25 intended business majors?

Mindy: Yeah.

Claire: And why did you choose 15 to 25?

Mindy: Just because you can see there are way more here (motions to outcomes near the center), so that gives you probably a higher probability of getting between that. Even with one sample, I think you'd probably get between 15 and 25.

Mindy said she thought that one sample could result in any outcome ranging from 15 to 25, indicating she viewed the collection of 500 sample outcomes as possible outcomes for one random sample of 100 undergraduates from North University. When I asked what changed in her thought process from the first range she gave (10 to 30) to the updated range (15 to 25), she explained, "I can see more random samples...more of them are between 15 to 25. So even if I just did one, I have a higher probability of getting between those numbers." Mindy repeatedly mentioned drawing only one sample: "If you just get one random sample," "I think with one it could be," "Even with one sample," "Even if I just did one." This is evidence that Mindy understood that the one random sample drawn from North University is simply one possible sample out of the 500 samples she simulated using the web-based applet. Furthermore, Mindy

assessed the likelihood of many possible outcomes when she said that one random sample had a “higher probability” of being in the range 15 to 25. All of these aspects of Mindy’s reasoning provide strong evidence that she understood that one random sample is simply one case out of many similar cases, for which probabilities can be estimated—a critical understanding when reasoning about sampling distributions.

In this section, I discussed my analysis of seven students’ reasoning throughout Tasks B and C based on four components. Although I did not give examples of all four components for all seven students, the examples I provided were typical of the group. Thus, imagining a hypothetical student who reasons in the ways I described for each of the four components illustrates a typical student’s reasoning in this group of seven. Recall, each of these seven students reasoned both inductively and deductively to continually adjust and refine what sample outcomes they expected and found surprising, supporting their understanding of how far the sample outcomes vary from the population parameter. In other words, these students were able to provide a reasonable estimate for sampling variability. In addition, they showed evidence of viewing data as an aggregate (Component 1 or Component 2), understanding how far the sample outcomes should vary from 20% (Component 3), and understanding that the collection of the 500 sample outcomes they drew using the web-based applet represents possible outcomes for one random sample of 100 undergraduates from North University—all critical aspects of reasoning about sampling distributions.

Providing an Unreasonable Updated Range of Values. Of the 10 students who repeatedly reasoned inductively and deductively to adjust their estimates for sampling variability, three included at most one of the four components in their reasoning throughout Tasks B and C (see Table 5). An important distinction between these three students’ reasoning and the other

seven students' reasoning is that these three students provided an updated range of values that was either unreasonable or not centered at 20% *and* they justified their range primarily on the collections of sample outcomes they drew from the box of beads, rather than the 500 random samples they drew using the web-based applet. In other words, none of these three students included Components 3 or 4 in their reasoning. Thus, these three students showed an underdeveloped understanding of sampling distribution.

Becky was the only student of the three who showed evidence of one of the four components. When comparing her predicted and observed sample outcomes, she indicated her predictions were “more spread out” than her observed sample outcomes for both the set of five sample outcomes and 20 sample outcomes, suggesting some evidence of an aggregate view of data because she compared the spreads of the two distributions (Component 1), a characteristic that describes the data as a whole. Furthermore, in her comparison for the 20 samples, she said, “The [observed] sample proportions were pretty representative of the population's true proportion of 20% ...the actual observations fall really close to the true mean proportion of 20%.” Here, Becky identified the center of the distribution of observed outcomes as 20% and described the variability from the center when she said the actual outcomes were “pretty representative of” and “fall really close to” the true parameter. Although she did not explicitly compare these measures for both distributions, her identification of the center as 20% and her description of the variability from the center show that she attended to summary characteristics of a distribution (i.e., center and variability from center) to describe the entire collection of outcomes. In contrast, both Julie and Corrina compared individual sample outcomes when they compared the distributions of their predicted and observed outcomes for both the set of five sample outcomes and 20 sample outcomes. For example, in her comparison for 20 sample

outcomes, Julie said, “They're somewhat similar, but I had no 20 percentages when I predicted a large amount.” Here, Julie compared the frequency of an individual outcome. When I asked her to explain what she meant by “somewhat similar,” she replied,

Somewhat similar in the aspect that I knew some would be below and above, and the highest that I did, my predicted was 27% and I did get 27%. The lowest I predicted was 10%, but I didn't get that. The lowest I got was 15%. So that part probably wasn't as accurate, but I had a few that were at the 23%, 22, 21, 25, and then 18 and 17. So I did see that in my actual results.

In this excerpt, Julie compared the minimum values, the maximum values, and the frequencies of several individual outcomes, rather than summary characteristics such as shape, center, and spread. Interestingly, Julie never looked at the parallel dot plots when she compared her 20 predictions to her 20 observed sample outcomes. Instead, she looked at the table of values for both. When Julie compared her 5 predictions to her 5 observed sample outcomes, she briefly looked at the parallel dot plots, then shuffled through the task papers to find where she recorded her predicted and observed outcomes and looked at these tables when she described how the outcomes compared. This was surprising because all other students—even Corrina who also compared individual sample outcomes—looked at (and many pointed at various aspects of) the parallel dot plots. Because both Julie and Corrina compared individual outcomes and never identified or compared characteristics like shape, center, or spread, they did not show evidence of identifying patterns in a distribution, suggesting a *local understanding* of data (Ben-Zvi, 2004) rather than a global view. Furthermore, neither assessed nor compared the probabilities of one or more sample outcomes (Component 2), showing additional evidence they did not view data as an aggregate.

From Table 5, these three students provided an updated range of values at the end of Task C that was either unreasonable or not centered at 20%. More importantly, their justifications for their ranges primarily focused on the outcomes from the physical samples they drew from the box of beads, rather than the 500 random samples they drew using the applet, or the outcomes of all the samples they drew throughout Tasks B and C, both from the box of beads and using the applet (Component 4). This result indicates that these three students did not view the collection of 500 sample outcomes as the best empirical sampling distribution. In other words, they did not view the sampling distribution of the 500 sample outcomes from the applet as being a better representation of all possible outcomes than the sampling distribution of the 20 or five sample outcomes from the box of beads. In Julie's case, she thought that a smaller collection of sample outcomes, such as five or 20, was a better indication than a larger collection of sample outcomes, such as 500, for what she could expect to get in one random sample. This may be because, in the context of the problem, she was drawing only *one* random sample, and one is closer to five and 20 than it is to 500. At the beginning of Task A, Julie gave a range of 10 to 30; a range that is both reasonable and centered at 20%. At the end of Task C, Julie gave an updated range of 15 to 27; a range that is reasonable but not centered at 20%. By itself, this updated range does not provide much insight into how Julie reasoned about the repeated sampling process, but her justification for her updated range does. Prior to giving her range, Julie asked, "Am I only doing it one time?" and then provided her range of 15 to 27, without looking at the dot plot displaying the 500 sample outcomes. When she justified her range, she said,

So, when we had 500 samples, you had a good bit actually at 15, but you didn't have that many at 27. But I picked it because if you're only doing it one time, when the [numbers

of samples] were quite lower (looks at her table of 20 sample outcomes on the task sheets), we were seeing numbers at around that range.

The outcomes of Julie's 20 random samples that she drew from the box of beads ranged from 15 to 27. Thus, it is reasonable to think that Julie chose her updated range based on these 20 sample outcomes, especially because she did not look at the dot plot of the 500 sample outcomes when she gave the updated range. In her justification, she talked about what she noticed in the 500 sample outcomes with respect to her bounds of 15 and 27, but she specifically said that sample outcomes for the smaller numbers of samples (e.g., 20) were around 15 to 27. When I asked what changed in her thinking from the first range she gave to her updated range, Julie again talked about the number of samples she drew. She said,

Well, from before I was expecting to see some at 10, but when we were doing our trials and then based on this one where we did 500, there rarely was ones that were as low as 10. So that's why I shifted it to higher because we still did see a good bit at 15...I gave 27 based on the five trials that we did and the 20 trials, which is why I changed that. But based on the graph that we have now for the 500, 27 would be quite high, but I changed it to 27 because we're only doing one, and this was with 500 samples.

In this excerpt, Julie tried to mediate what she noticed in the results of the smaller number of samples with what she noticed in the results of the 500 samples. For example, she justified her upper bound of 27 based on her previous results of her five and 20 samples but indicated that a sample outcome of 27 was "quite high" in the results of the 500 samples, but she ultimately chose 27 because "we're only doing one, and this was with 500 samples." Julie's repeated focus on drawing only one sample, coupled with her justification based on the smaller number of samples rather than the 500 samples, shows evidence that she thought the outcomes of a smaller

collection of samples was a better indicator for what to expect in one random sample. This approach indicates that Julie did not view the collection of 500 sample outcomes as the best approximation for the theoretical sampling distribution, from which she could estimate probabilities.

Becky and Corrina did not discuss the number of samples drawn in their justifications for their updated ranges. Instead, they referenced what they noticed in their previous sample outcomes, without looking at or mentioning the results of their 500 samples outcomes. For example, Corrina said, “I guess I would expect 25 to 17 again...just from earlier when we did the scooping thing, that's kind of what it was around.” Becky noticed her previous sample outcomes “were falling pretty much between 11 to 25 ish” and gave a range of 5 to 32 to “allow some room” for lower and higher numbers. Becky and Corrina did not show the same reasoning as Julie about the number of samples. Furthermore, Becky and Corrina’s explanations tended to be difficult to follow because they were either convoluted or quite brief, even after additional prompting from me. Without additional evidence, my only hypothesis for why Becky and Corrina did not use the results of the 500 random samples in their justification is because they trusted the results of the samples they drew from the box of beads—samples they could see, touch, and count—more than the results of the samples they drew using the applet—samples that were constructed from computer programming. I will discuss this potential limitation of the applet, along with others, in the last chapter.

In this section, I discussed my analysis of three students’ reasoning throughout Tasks B and C based on four components. Recall, each of these students reasoned both inductively and deductively to adjust what sample outcomes they expected. Although they provided updated ranges that were not centered at 20% (Component 3) nor justified based on the results of the 500

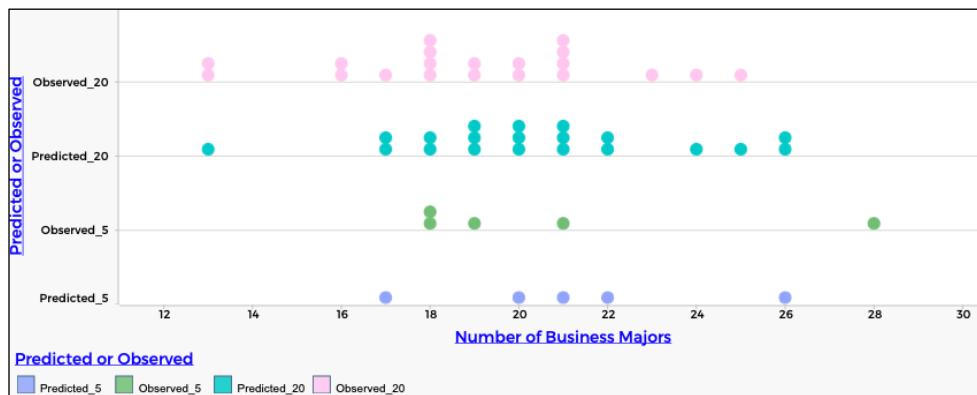
sample outcomes (Component 4), their updated ranges were more reasonable than their initial ranges, indicating they were able to provide a more reasonable estimate for sampling variability compared to their original estimate. Only Becky showed some evidence of viewing data as an aggregate when she compared her predicted and observed outcomes based on summary characteristics of a distribution (Component 1), though none assessed or compared the likelihood of any sample outcomes (Component 2).

Tyra's Reasoning: Using Deductive Reasoning to Identify Expected and Surprising Outcomes

Returning to the broader forms of reasoning students used in Tasks B and C, Tyra was the only one who did not use the results of her previously drawn samples to inform her thinking about later sample outcomes. In Task A and prior to drawing any samples, Tyra gave a range of 15% to 25% for the percent of intended business majors she would expect to get in a random sample of 100 undergraduates from North University. Like several others, she chose her 5% margin based on a personal feeling. She said, "Just because five below, five above...if I wanted it to be closer, I could have done three below, three above, but I just wanted a bigger gap, I guess more room for error, just a random number I picked." Unlike others, she did not use the results of any of her observed outcomes to inform her thinking about later collections of samples; she did not reason inductively. Instead, she repeatedly reasoned deductively by applying previously identified rules to new sample outcomes. For example, after predicting and drawing 20 random samples, I asked Tyra if she would be surprised by sample outcomes of 24% and 38%. Figure 15 displays the parallel dot plots of Tyra's predicted and observed outcomes for 20 random samples. I expected Tyra to use the results of the random samples she had drawn from the box of beads to assess the likelihood of each of these outcomes. For instance, I expected Tyra might induce that 38% is a very unlikely outcome because, out of the 20 (or 25 total) samples

Figure 15

Tyra's Predicted and Observed Sample Outcomes



she drew, none produced outcomes higher than 28. Instead, Tyra justified whether she would be surprised by these sample outcomes based on how close each were to 20% without looking at or talking about her observed sample outcomes. Previously in Task A, Tyra quantified what she meant by “really close to 20” by giving a range of 15 to 25. It is reasonable that Tyra used her definition of “close”—her range of 15 to 25—to determine whether she was surprised by sample outcomes of 24% and 38%, especially because she did not look at or talk about the dot plot or table displaying her observed sample outcomes. Tyra said she would not be surprised by a sample outcome of 24% because “24 is super close to 20, and just because four extra people happen to be walking around isn't that surprising.” Similarly, Tyra’s decision about whether 38% was a surprising outcome was based on its “closeness” to 20%. She said, “I would be a little surprised just because 38 is a lot bigger than 20,” and added, “But not like, oh my God, that should not happen. It'd be like, oh, that's a little out of the ordinary, but nothing that's going to ruin my, whatever I'm doing with the sample.” Instead of assessing the likelihood of getting a sample outcome of 38% based on her 20 sample outcomes, Tyra again used her definition of “close” to determine that 38% was a surprising outcome. For each outcome, Tyra applied her

definition of “close” (rule) to a sample outcome (case) to deduce whether the outcome was surprising (result).

Regarding the four components I identified in my analysis of the other 10 students’ reasoning (see Table 5), Tyra only included Component 3 in her reasoning when she provided an updated range of 15 to 25; a range that is both reasonable and centered at 20%. Of course, the absence of Component 4 was a direct result of Tyra never having used the outcomes of her previously drawn samples to inform her thinking about future samples outcomes. In other words, because Tyra never reasoned inductively in Tasks B and C, she could not have justified her updated range of values based on the results of the 500 samples (Component 4). In addition, Tyra did not show evidence of viewing data as an aggregate because she did not compare her predicted and observed outcomes using characteristics such as shape, center, and spread (Component 1) nor did she assess or compare the probabilities of one or more sample outcomes (Component 2). Instead, Tyra compared individual sample outcomes. For example, Tyra identified discrepancies in her five predictions by comparing each one to her observed sample outcomes. She said,

I put one at 26% and got one at 28%, so I thought there'd be one little higher and there was. I put one right at 20%, I didn't get 20%, but I got 19 and 21. I did put 21 and got 21, and then I put 22 and didn't get a 22.

She compared her predicted and observed outcomes for the 20 random samples in a similar way. Because she focused on individual outcomes and did not identify or compare patterns in the distributions, Tyra did not show evidence of viewing the collections of sample outcomes as a whole, suggesting a local understanding of data (Ben-Zvi, 2004).

Recall, at the end of Task A, Tyra said she would expect to get 15 to 25 intended business majors in a random sample of 100 undergraduates from North University and justified her margin based on a personal feeling. At the end of Task C, Tyra gave the same range of values. By itself, this is not surprising. In fact, Lyla also gave the same range at the beginning and end of Interview 1. What was the most surprising aspect of Tyra's reasoning in Tasks B and C was that she justified her same range with the same reason—a personal feeling. I asked Tyra what changed in her thinking from the first time she gave the range to the second time. Without looking at the dot plot of the 500 sample outcomes, she said,

I think five is still kind of my reasoning, but also, it's still close to 20. I don't know, five is just a good number. Five. I don't know. I could have used six, but I didn't. It's still pretty close to it. And so yeah, I feel like five below and five above. I wanted to use the same number below and above.

Although Tyra's range is reasonable and centered at 20%, it is unclear what she understands about how far the sample outcomes vary from 20%. Recall, when she gave her range initially in Task A, Tyra said, "I could have done three below, three above, but I just wanted a bigger gap" and identified five as "just a random number" she picked. When she gave the same range at the end of Task C, she said, "Five is a good number...I could have used six, but I didn't." Both of Tyra's explanations for choice of a 5% margin suggest that she chose 5% arbitrarily, especially because she provided two examples of other values she could have chosen for the margin.

It was extremely surprising that Tyra never used the results of any of her sample outcomes to inform her thinking. In the overview, I provided two possible explanations for this. One possible explanation is that the results of Tyra's samples did inform her thinking, but her words and actions did not convey that she used the results of her samples in her reasoning.

Another explanation is that Tyra did not view the box of beads or the web-based applet as models of the North University context. However, Tyra's detailed explanation of how the different aspects of the box of beads related to the North University context suggests otherwise. Furthermore, in Task C, Tyra identified how each of the parameters in web-based applet related to the North University context. Based on my analysis, there is no more evidence for one explanation than the other.

Summary

In this chapter, I discussed the results of my analysis of all 11 students' reasoning in Interview 1. In Tasks A, B, and C, students repeatedly reasoned about what sample outcomes they would expect and find surprising from a population with a known parameter. Recall, I aimed to answer Research Question 1 with the results of Interview 1. In Research Question 1, I asked:

When given a population with a known parameter, what forms of reasoning do novice statistics students employ when determining what sample outcomes they expect and what sample outcomes they find surprising, and what do these forms of reasoning reveal about their reasoning *about* sampling distributions?

When given a population with a known parameter, students tended to reason differently prior to and after drawing samples to determine what sample outcomes they expected and found surprising; they tended to reason deductively prior to drawing samples and inductively after drawing samples. Table 8 displays a visual model of how students shifted between deductive and inductive reasoning based on the task. A pink rectangle represents a reasoning clip, or a student's response to a prompt in which they included an explanation or a justification; a blue rectangle

indicates where a student reasoned deductively; and a green rectangle indicates where a student reasoned inductively. I identified five reasoning clips in Task A, represented by the five pink

Table 8

A Visual Model of Students' Shifts in Forms of Reasoning in Interview 1

Student	Task A		Task B			Task C		
	Drew 0 samples		Drew 5 samples		Drew 20 samples		Drew 500 samples	
Tami	<div><div></div><div></div><div></div><div></div><div></div><div></div></div> <div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div></div> <div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div></div> <div><div></div><div></div></div>	<div><div></div></div> <div><div></div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div><div></div></div> <div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div></div> <div><div></div><div></div><div></div></div>	
Julie	<div><div></div><div></div><div></div><div></div><div></div><div></div></div> <div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div></div> <div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div></div> <div><div></div><div></div><div></div></div> <div><div></div></div>	<div><div></div></div> <div><div></div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div><div></div></div> <div><div></div></div> <div><div></div><div></div></div>	<div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div></div> <div><div></div><div></div></div> <div><div></div></div>	
Jess	<div><div></div><div></div><div></div><div></div><div></div><div></div></div> <div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div></div> <div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div></div> <div><div></div><div></div><div></div></div> <div><div></div></div>	<div><div></div></div> <div><div></div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div><div></div></div> <div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div></div> <div><div></div><div></div></div> <div><div></div></div>	
Becky	<div><div></div><div></div><div></div><div></div><div></div><div></div></div> <div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div></div> <div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div></div> <div><div></div><div></div><div></div></div> <div><div></div></div>	<div><div></div></div> <div><div></div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div><div></div></div> <div><div></div></div> <div><div></div><div></div></div>	<div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div></div> <div><div></div><div></div></div> <div><div></div></div>	
Lorraine	<div><div></div><div></div><div></div><div></div><div></div><div></div></div> <div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div></div> <div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div></div> <div><div></div><div></div><div></div></div> <div><div></div></div>	<div><div></div></div> <div><div></div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div><div></div></div> <div><div></div><div></div></div> <div><div></div><div></div></div>	<div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div></div> <div><div></div><div></div></div> <div><div></div></div>	
Tyra	<div><div></div><div></div><div></div><div></div><div></div><div></div></div> <div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div></div> <div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div></div> <div><div></div><div></div><div></div></div> <div><div></div></div>	<div><div></div></div> <div><div></div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div><div></div></div> <div><div></div><div></div></div> <div><div></div><div></div></div>	<div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div></div> <div><div></div><div></div></div> <div><div></div></div>	
Waverly	<div><div></div><div></div><div></div><div></div><div></div><div></div></div> <div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div></div> <div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div></div> <div><div></div><div></div><div></div></div> <div><div></div></div>	<div><div></div></div> <div><div></div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div><div></div></div> <div><div></div><div></div></div> <div><div></div><div></div></div>	<div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div></div> <div><div></div><div></div></div> <div><div></div></div>	
Eric	<div><div></div><div></div><div></div><div></div><div></div><div></div></div> <div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div></div> <div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div></div> <div><div></div><div></div><div></div></div> <div><div></div></div>	<div><div></div></div> <div><div></div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div><div></div></div> <div><div></div><div></div></div> <div><div></div><div></div></div>	<div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div></div> <div><div></div><div></div></div> <div><div></div></div>	
Lyla	<div><div></div><div></div><div></div><div></div><div></div><div></div></div> <div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div></div> <div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div></div> <div><div></div><div></div><div></div></div> <div><div></div></div>	<div><div></div></div> <div><div></div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div><div></div></div> <div><div></div><div></div></div> <div><div></div><div></div></div>	<div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div></div> <div><div></div><div></div></div> <div><div></div></div>	
Mindy	<div><div></div><div></div><div></div><div></div><div></div><div></div></div> <div><div></div><div></div><div></div><div></div><div></div><div></div></div>	<div><div></div></div> <div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div></div> <div><div></div><div></div><div></div></div> <div><div></div></div>	<div><div></div></div> <div><div></div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div><div></div></div> <div><div></div><div></div></div> <div><div></div><div></div></div>	<div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div></div> <div><div></div><div></div></div> <div><div></div></div>	
Corrina	<div><div></div><div></div><div></div><div></div><div></div><div></div></div> <div><div></div><div></div><div></div><div></div></div>	<div><div></div></div> <div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div></div> <div><div></div><div></div><div></div></div> <div><div></div></div>	<div><div></div></div> <div><div></div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div><div></div></div> <div><div></div><div></div></div> <div><div></div><div></div></div>	<div><div></div><div></div></div> <div><div></div></div>	<div><div></div><div></div><div></div><div></div><div></div><div></div></div> <div><div></div><div></div></div> <div><div></div></div>	
Key	<div></div>	Reasoning Clip	<div></div>	Deductive Reasoning	<div></div>	Inductive Reasoning		

rectangles in the first column of Table 8, eight reasoning clips in Task B, and five reasoning clips in Task C. Prior to drawing samples, I expected that students would reason based on informal facts or rules recalled from their recently completed introductory statistics course; after drawing samples, I expected that students would reason based on the results of their samples. Table 8 shows that the results of Interview 1 partially aligned with my expectations. However, the different forms of reasoning revealed important distinctions in students' understanding of sampling distributions. Recall, students did not draw any samples in Task A, drew five then 20 samples from the box of beads in Task B, and drew 500 samples using the web-based applet in Task C. Table 8 shows that all 11 students reasoned deductively in Task A. All 11 students came into the study with knowledge of the *existence* of sampling variability. However, not all students came to the study with the same intuition about *how far* the outcome of a random sample varies from the population parameter. At the end of Task A, students provided a range of sample outcomes they would expect to get in a random sample of 100 undergraduates from North University. Students' response to this prompt was coded in the fifth reasoning clip, represented by the fifth pink rectangle in the first column of Table 8. Students who reasoned deductively provided a range of sample outcomes that was both reasonable and centered at the population parameter; those who did not reason deductively—Becky, Waverly, and Corrina—provided a range that was both unreasonably wide and not centered at the population parameter.

After drawing samples, students tended to use a combination of inductive and deductive reasoning to determine what sample outcomes they expected and found surprising. Table 8 shows that all students except Tyra reasoned inductively two or more times, each time after having drawn some number of random samples either from the box of beads or using the web-based applet. In Tasks B and C, those who reasoned inductively used the results of the samples

they had drawn to repeatedly adjust and refine their estimate for how far the sample outcome should vary from the population parameter. Although most students who provided an initial range in Task A that was reasonable also provided an updated range in Task C that was reasonable, they developed a more robust understanding of sampling variability from identifying patterns in their observed sample outcomes. For each student, Table 9 displays the initial range, the final range, and the form of reasoning they used in their rationales. Inductive reasoning was particularly powerful for Becky, Waverly, and Corrina, who initially provided unreasonably wide ranges. By the end of Task C, these students provided a narrower range to estimate how far the sample outcomes should vary from the population parameter.

Table 9

Students' Reasoning Corresponding to Their Initial and Final Range of Values

Student	Initial Range	Form of Reasoning	Final Range	Form of Reasoning
Tami	15 to 25	Deductive; margin based on personal feeling	16 to 24	Inductive; based on 500 samples outcomes
Jess	10 to 30	Deductive; margin based on personal feeling	15 to 25	Inductive; based on all sample outcomes
Lorraine	13 to 27	Deductive; margin based on sample size	11 to 29	Inductive; based on 500 sample outcomes
Eric	15 to 25	Deductive; margin based on previous knowledge	10 to 30	Inductive; based on 500 sample outcomes
Lyla	10 to 30	Deductive; margin based on personal feeling	10 to 30	Inductive; based on all sample outcomes
Mindy	10 to 30	Deductive; margin based on personal feeling	15 to 25	Inductive; based on 500 sample outcomes
Waverly	10 to 50		10 to 30	Inductive; based on 500 sample outcomes
Julie	10 to 30	Deductive; margin based on sample size	15 to 27	Inductive; based on 20 sample outcomes
Becky	0 to 80		5 to 32	Inductive; based on 20 sample outcomes
Corrina	5 to 40		17 to 25	Inductive; based on 20 sample outcomes
Tyra	15 to 25	Deductive; margin based on personal feeling	15 to 25	Deductive; margin based on personal feeling

Thus, inductive reasoning was useful in helping these students develop a better understanding of sampling variability. Tyra was the only student who never used the results of her sample outcomes to inform her thinking and showed no evidence that her understanding of sampling variability changed from her initial intuition at the beginning of Task A.

Although inductive reasoning was useful in helping these students develop a better understanding of sampling variability, this form of reasoning does not provide a complete model of a student's understanding of sampling distribution. For example, although inductive reasoning was necessary for Becky to provide a better estimate for sampling variability than her initial range, other aspects of her reasoning (i.e., the four components) revealed an underdeveloped understanding about sampling distributions.

CHAPTER 6

REASONING WITH SAMPLING DISTRIBUTIONS

The goal of Interview 2 was to examine how novice statistics students reasoned under uncertainty when given a population with an unknown population parameter. Specifically, I aimed to understand how students reasoned from one sample outcome to draw conclusions about a larger population. Students engaged with only one task, Task D, in Interview 2. The context of Task D was like the previous three tasks in that it was about the percentage of intended business majors at a large university. However, Task D differed in several ways: (1) the context involved a different university, South University; (2) the proportion of all undergraduates at South University who intend to major in business was unknown; (3) students began with information about a single sample they drew from a box of beads whose contents were hidden from them; and (4) students predicted the unknown population parameter. Throughout Task D, students used the web-based applet to predict and test multiple values for the unknown population parameter and ultimately provided an estimate for the true percentage of intended business majors at South University. Thus, the results of Interview 2 help answer Research Question 2. Recall, in Research Question 2, I asked:

When given a population with an unknown parameter, what forms of reasoning do novice statistics students employ when making inferences from one sample outcome to the population from which it was drawn, and what do these forms of reasoning reveal about their reasoning *with* sampling distributions?

Surprisingly, not all students used the result of their one sample to draw conclusions about the population from which it was drawn. In other words, not all students made inferences *from sample data* to a larger population with an unknown parameter of interest. Students who used the result of one sample reasoned abductively or deductively to determine a range of plausible values for the unknown parameter. Students who did not use the result of one sample outcome either (a) reasoned deductively or inductively, or (b) did not show evidence of any of the three broader forms of reasoning when determining their range of values for the unknown population parameter.

In this chapter, I first give a description of Task D by listing specific prompts and activities in which students engaged. Next, I provide an overview of the major findings, describing differences in how students used the result of their sample outcome and identifying the broader forms of reasoning within each group. Then I support the major findings by describing and comparing the reasoning of two or more representative students within each of the forms of reasoning. Throughout these sections, I identify what the broader forms of reasoning and the more nuanced differences in reasoning revealed across the student pool.

Results from Task D

Description of Task D

At the beginning of Interview 2, I introduced students to another large university, South University. The introduction in Task D read,

Recall from the first interview, North University reported on their website that 20% of the undergraduate students are intended business majors. A nearby large university, South University, does *not* report on their website what percentage of their undergraduate students are intended business majors. You know it is impossible to ask every single

undergraduate student at South University what their intended major is, so you decide to ask a random sample of undergraduates instead.

Next, I explained that we could model the context with a box of plastic beads, noting that the box was different from the one they used in the first interview in that we did not know the proportion of beads in each color category. Recall, the box of beads was covered so that students could not see the contents of the box (see Figure 16). In Task D, students first drew one random sample of 100 beads from the box and explained how they thought that their sample related to the population of all undergraduate students at South University. After they determined if they thought the percentage of all undergraduate students at South University that were intended business majors was the same as North University's 20%, I asked them to make a prediction for South University's true percentage based on their sample. Next, I asked students to explain how they could determine if the outcome of their one random sample drawn from the box of beads indicated that the true parameter for South University was equal to their initial prediction.

Figure 16

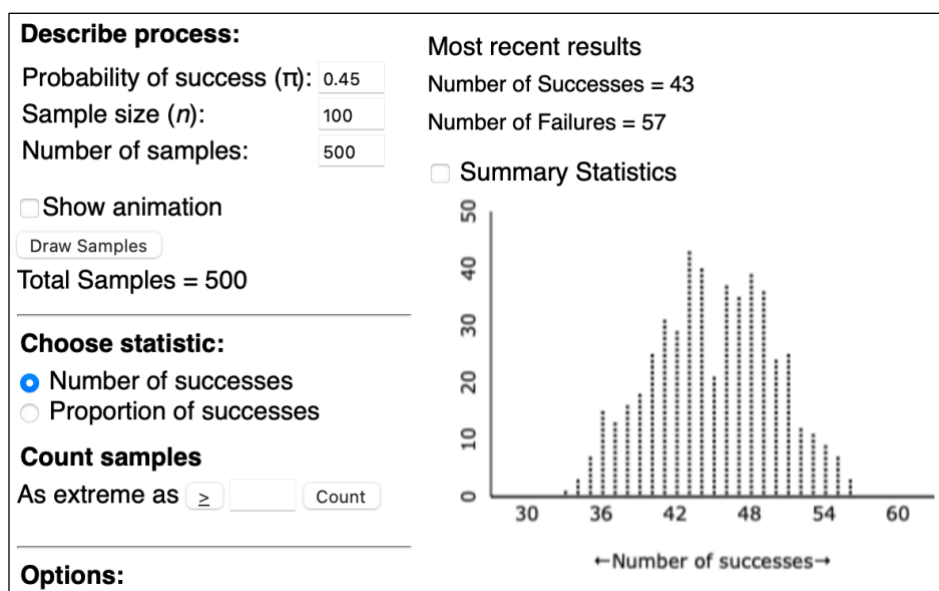
Box of Beads From Which Students Drew One Sample in Task D



Students then used the web-based applet to test their prediction. I hoped that students would test their prediction using the web-based applet in the following way: Suppose the one random sample of 100 beads that I drew from the covered box that represents the population of undergraduate students at South University resulted in 42% yellow beads, or 42 out of 100 business majors. I predict the true percentage of undergraduates at South University who are intended business majors is 45% because this value is close to the outcome of my sample. To test my prediction of 45%, I set the probability of success in the applet to 0.45 and draw 500 random samples of 100 each. This means that I am telling the applet to draw 500 random samples of size 100 from a population whose true parameter is 0.45, or 45%. Alternatively, I am telling the applet to draw 500 random samples of 100 undergraduates from South University, *if* the true proportion of intended business majors at South University is 0.45, or 45%. Figure 17 shows a screenshot from the web-based applet after completing this process.

Figure 17

Testing a Prediction of 45% Using the Web-based Applet



The dot plot shows a collection of 500 samples outcomes; each dot represents the number of successes in a sample of 100 or, in context, the number of intended business majors in a sample of 100 undergraduate students; the vertical axis is the frequency of each outcome. Furthermore, the applet provides the distribution of the most recent sample. In this case, the 500th sample resulted in 43 success and 57 failures, or 43 intended business majors and 57 non-intended business majors. From the collection of 500 sample outcomes, my sample outcome of 42 (or one more extreme) occurred quite often in a population whose true parameter is 45%. In other words, *if* the true percentage of undergraduates at South University who are intended business majors is 45%, drawing a random sample of 100 undergraduates and getting a sample outcome like my 42% (or one more extreme) is probable. Therefore, I decide that 45% is a plausible value for the true percentage of undergraduates at South University who are intended business majors. After testing their initial prediction, I repeatedly asked students if there were other values they thought could be the true percentage; they continued to make and test several predictions for the unknown population parameter, and ultimately provided a range of what they considered to be plausible values for the true percentage of intended business majors at South University.

Overview of Findings From Task D

I hypothesized that reasoning abductively would be a productive way to reason with sampling distributions to draw inferences from sample data to a population with an unknown parameter because abductive reasoning shares aspects of inferential reasoning. Recall, abductive reasoning is the inference of a case from a rule and result: We start with an initial observation, or result. This result could be explained by the supposition that it is a case of some general rule. Therefore, we adopt that supposition (Peirce, 1878). Similarly, in statistical inference, we start with the outcome of a sample. By considering the probability of the sample outcome under

multiple hypothetical parameters, we conclude this sample outcome could be explained by the supposition that it was drawn from a population whose parameter is equal to p . Therefore, we adopt that supposition. In short, we start with a sample outcome (result) and, using probability, conclude that the sample could have been drawn from (is a case of) a population whose parameter is equal to p (general rule). In Task D, only three students—Lorraine, Eric, and Lyla—reasoned abductively in this way. All three continually made and tested predictions for the true percentage of undergraduates at South University who intended to major in business, assessing the plausibility of their predictions based on the likelihood of their sample outcome under multiple hypothetical parameters. Furthermore, they showed evidence of coordinating population distribution and sampling distribution by relating one or more of the sampling distributions they constructed using the web-based applet to a hypothetical population from which a collection of samples was drawn. These two critical components in their abductive reasoning showed evidence of a sophisticated understanding of sampling distribution and making inferences from sample data. Thus, abductive reasoning was essential in reasoning with sampling distributions to make inferences from sample data to a population with an unknown parameter.

Although four other students assessed the plausibility of their predictions based on the likelihood of their sample outcome, they did not show evidence of coordinating population distribution and sampling distribution. Instead, they used a rule—based on the likelihood of their sample outcome—to deduce which of the values that they tested should be included in their range of plausible values for the unknown parameter. They showed evidence of understanding that each sampling distribution they constructed represented a collection of sample outcomes, but they did not connect the collection of samples to the hypothetical population from which the

samples were drawn—a critical understanding to reason productively from sample data to the population from which it was drawn.

The remaining four students did not use the result of their sample outcome to determine their range of plausible values for the unknown population parameter. Although all four students understood that random samples will produce outcomes that vary from the population parameter (from Interview 1 results), they did not show evidence of applying their understanding of sampling variability to determine the likelihood of their sample outcome under multiple hypothesized values for the percentage of undergraduates at South University who are intended business majors.

In my analysis of the students' reasoning in Tasks D, I identified four components that were important in understanding how they reasoned with sampling distributions:

- (1) Students related one or more sampling distributions to a hypothetical population;
- (2) Students assessed the plausibility of one or more values for the unknown parameter by considering the likelihood of one sample outcome;
- (3) Students provided a range of values for the unknown population parameter based on the outcome of multiple random samples;
- (4) Students assessed the plausibility of multiple values for the unknown parameter by considering visual characteristics of the graph (e.g., shape, highest peak) of one or more sampling distributions.

Table 10 displays the categories of reasoning I identified for Task D. Within each form of reasoning, I identified which components each student included in their reasoning. I organized the results of Interview 2 based on whether students used the result of their one sample outcome to determine a range of plausible values for the unknown population parameter (see Table 10).

Table 10*Categories of Reasoning in Task D*

Reasoned From One Sample Outcome	Form of Reasoning	Component	Student										
			Tami	Julie	Jess	Becky	Lorraine	Tyra	Waverly	Eric	Lyla	Mindy	Corrina
Yes	Abductive	Related one or more sampling distributions to a hypothetical population					X			X	X		
		Assessed the plausibility of multiple values for the unknown parameter by considering the likelihood of one sample outcome					X			X	X		
	Deductive	Assessed the plausibility of multiple values for the unknown parameter by considering the likelihood of one sample outcome	X					X				X	X
No	Inductive	Provided a range of values for the unknown population parameter based on the outcome of multiple random samples			X								
	Deductive	Assessed the plausibility of multiple values for the unknown parameter by considering visual characteristics of the graph (e.g., shape, highest peak) of one or more sampling distributions		X									
	No evidence of the three forms of reasoning					X			X				

In the first section, I highlight the subtle yet critical distinction between abductive and deductive reasoning for those seven students who used the result of their sample outcome by describing and comparing the reasoning of two representative students for each broader form of reasoning. In the second section, I describe and compare the reasoning of the four students who did not use the result of their sample outcome to determine a range of plausible values for the unknown parameter. Throughout both sections, I support the major findings by identifying what the broader forms of reasoning and the more nuanced differences in reasoning revealed across the student pool.

Drawing Inferences About a Population Using the Result of One Random Sample

From Table 10, seven students used the outcome of the one random sample they drew from the box of beads to determine a range of values they considered to be plausible for the true percentage of undergraduates at South University who were intended business majors. Specifically, these seven students used the likelihood of their sample outcome to assess the plausibility of multiple values for the population parameter (Component 2). However, the form of reasoning the students employed revealed a subtle yet critical component of their understanding of sampling distribution. Those who reasoned abductively showed evidence of coordinating at least two levels of distribution—population distribution and sampling distribution (Component 1)—indicating they understood that each sampling distribution they constructed represented a collection of sample outcomes that could have been produced by some hypothetical population. Those who reasoned deductively showed evidence of understanding that each sampling distribution represented a collection of sample outcomes, but they did not relate the collection to the population that could have produced the collection—a critical component in drawing inferences from sample data to a population with an unknown parameter.

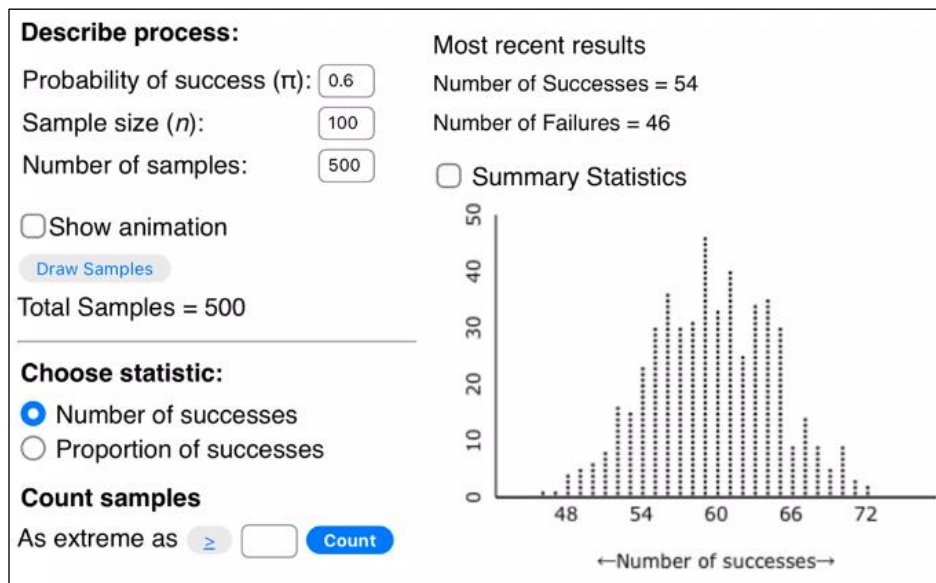
Abducting Plausible Values for the Unknown Population Parameter. Only three students—Lorraine, Eric, and Lyla—related one or more sampling distributions they constructed using the applet to a hypothetical population that could have produced the collection of samples outcomes (Component 1). These students abducted a range of values based on the likelihood of the outcome of their one random sample that they drew from the box of beads that modeled the South University context. They repeatedly predicted and tested multiple values for the unknown parameter and assessed the plausibility of each value by considering how often their sample occurred in the many samples they drew using the web-based applet (Component 2). Each time they determined that a hypothesized value was plausible, they abducted a case from a result and a rule. Starting with their one sample outcome (result), they hypothesized a value for the unknown population parameter, then constructed a sampling distribution in the web-based applet with the probability of success equal to the hypothesized value and drew many samples. Because their sample outcome was likely to have occurred under the hypothesized value (rule), they abducted that their sample could have been drawn from a population whose parameter was equal to the hypothesized value (case).

For example, when testing her prediction of 60% for the unknown population parameter, Lorraine drew 500 random samples with the web-based applet using a probability of success of 0.6. Figure 18 shows the results of her simulation. Using the results of her simulation, Lorraine assessed the likelihood of her sample outcome of 52% (Component 2) when the probability of success was 0.6. She said,

It looks like 52 would be reasonable...it looks like there's a pretty significant number of results that are...52. And then past that is where it tends to kind of taper out. So maybe if the true proportion were 60%, the lower bound would be like 48 or so.

Figure 18

The Result of Lorraine's Simulation When Testing 60%



Given her sample outcome of 52% (result), Lorraine assessed the likelihood of her sample outcome, determining that her sample outcome “would be reasonable” “if the true proportion were 60%” (rule). She abduced that her sample outcome could have come from a population whose true parameter was 60% (case) and included 60% in her range of values for South University’s true parameter. Lorraine related the sampling distribution she constructed to a hypothetical population (Component 1) with a “true proportion” of 60%, indicating she understood that the sampling distribution represented the outcomes of 500 random samples drawn from a population whose parameter was 60%. Lorraine continued to use the applet in this way to abduce multiple values that were plausible for the true percentage of undergraduates at South University who were intended business majors, ultimately providing a range of 42% to 62% for the unknown population parameter.

Like Lorraine, Eric and Lyla repeatedly predicted and tested multiple values for the unknown population parameter, assessing the plausibility of each value based on how often their

sample outcome occurred (Component 2). Furthermore, both showed evidence of relating one or more sampling distributions they constructed using the applet to a hypothetical population that could have produced that collection of sample outcomes (Component 1). When Eric tested his prediction of 40%, he explained that the 40% represented “the assumption that if South University had their proportion online...that proportion is 40%. It’s like the 20% from North University,” indicating that he related the sampling distribution he constructed with a probability of success of 0.40 to a hypothetical population whose true parameter was equal to 40% (Component 1). To assess the plausibility of his prediction, he considered the likelihood of his sample outcome (Component 2) under this assumption. He said, “If you say it’s 40, then 50 could be a possibility,” pointed to a sample outcome of 50 on the dot plot, and said, “So we could say, oh, that one [sample] that I drew was that one.” When Lyla tested her prediction of 50%, she explained that she “made a distribution based on this 0.5” and looked for “the probability of drawing 46 students in a sample of 100 from this larger population.” She related the sampling distribution she constructed with a probability of success to a “larger population” (Component 1) and considered the likelihood of her sample outcome of 46 (Component 2). Like Lorraine, both Eric and Lyla repeatedly reasoned in this way, abducting a case from a result and a rule. They abducted that their sample outcome (result) could have been drawn from a population (case) whose parameter was equal to their prediction (rule).

The conclusion of this abductive inference is that the one random sample these students drew could have been drawn from a population with some parameter, p . Without connecting a collection of sample outcomes to the population from which they were drawn, this conclusion is not possible. Without this connection, considering the likelihood of the outcome of the one sample drawn from the box of beads (Component 2) can be used as a procedure to deduce a

range of values for the unknown population parameter. I hypothesize that this was the case for Tami, Tyra, Mindy, and Corrina.

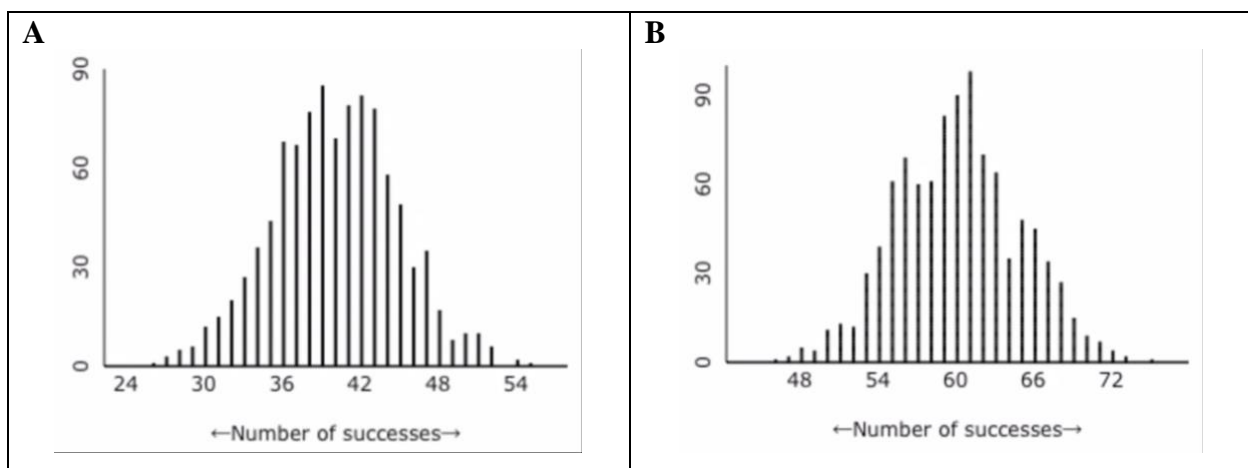
Deducing Plausible Values for the Unknown Population Parameter. Although Tami, Tyra, Mindy, and Corrina used the likelihood of their sample outcome to assess the plausibility of multiple values for the unknown parameter (Component 2), they did not relate each sampling distribution they constructed to a hypothetical population that could have produced those sample outcomes (Component 1). They did not show evidence of hypothesizing what population could have produced their sample outcome. Instead, they used the likelihood of their sample outcome as a rule to determine whether they should include their prediction in the range of values for the unknown population parameter. These students reasoned that if their sample outcome occurred often when the probability of success was equal to p , then p should be included in their range of values, without connecting the collection of sample outcomes to a population whose parameter is equal to p . This connection is the subtle yet critical distinction between the abductive inference I described in the previous section and these four students' deductive inference. By considering the likelihood of their sample outcome each time they constructed a new distribution using the applet (Component 2), they understood that the distribution they constructed represented a collection of sample outcomes to which they could compare their one sample outcome. However, they did not show evidence of understanding that each sampling distribution they constructed with some fixed parameter, p , represented a collection of possible outcomes from a population whose true parameter was equal to p .

Like Lorraine, Eric, and Lyla, these four students repeatedly assessed their predictions based on the probability of their sample outcome (Component 2). Corrina's reasoning is representative of the group. When testing her prediction of 40%, Corrina considered the

likelihood of her sample outcome of 51. She said, “It’s kind of on the higher end. It’s not as likely, but it totally could happen” and included 40% in her range at the end of Task D (see Figure 19a). When she tested 60%, she said, “Yeah, so that 51, again, is kind of on the lower end, but still likely, so I think it could be 60” (see Figure 19b) and, again, included 60% in her final range of 38% to 62% at the end of Task D. Corrina applied a general rule to each of her predictions to deduce whether she should include the prediction in her range of values for the true parameter at South University: If my sample outcome occurred in a large collection of outcomes when the probability of success is equal to p , then p should be included in my range of values. She applied her rule to her prediction of 40% and reasoned that because her sample outcome of 51 “could totally happen” when the probability of success was equal to 40% (case), then 40% was a plausible value for the unknown parameter (result). Similarly, because her sample outcome of 51 was “on the lower end, but still likely” when the probability of success was 60% (case), then 60% was a plausible value for the unknown parameter (result).

Figure 19

The Results of Corrina’s Simulations When Testing 40% and 60%



Students' Summaries. At the end Task D, the students summarized how they determined their range of plausible values for the true percentage of intended business majors at South University. These summaries provided additional evidence of the differences in the students' understanding across the two groups. Table 11 shows the summaries of a subset of students who are representative of each group. There are two important differences between the summaries of students who reasoned abductively and those of students who reasoned deductively. First, both Lorraine and Lyla mentioned the context of the problem multiple times when they referred to "business students," "business majors," "a sample of undergraduates," and "a sample of 100 students." Second, they both related the context to a hypothetical population when they referred to "the proportion of business students," "the true proportion of students that are intended business majors," and "students in a sample of 100 from this larger population." In addition to the explicit connections these students made to a hypothetical population in their explanations when testing multiple predictions, these key differences in their summaries provide additional evidence that this group of students coordinated two levels of distribution: population distribution and sampling distribution. In contrast, Corrina and Mindy referred to "the dot plot," "this teeny tiny dot," and "the graph," without ever referring to the context of the problem or using words or phrases that suggested they were thinking about a hypothetical population in relation to the samples they drew using the applet.

The subtle difference I identified in these two groups of students' reasoning was a critical component in my understanding of how they reasoned with sampling distributions to make inferences from sample data to a larger population. This result was surprising because, as a former statistics teacher, I would have thought that determining a range of values for an

Table 11*Students' Summaries in Interview 2*

Form of Reasoning	Summary
Abductive	<p>Lorraine: If I put in the probability of success as 42, or if I put 0.42 in the center, that would be the proportion of business students. Then it'll tell me, or it won't tell me, but I can guess how likely it is that I would pull the result that I pulled (motions to her sample of 100 beads on the scooper) from that [box of beads]...If it was likely enough that I could pull a sample of 100 students where 0.52 was 52 is the percent that were business majors, then it would be likely that that number could be within the range of the actual proportion...It's kind of a messy way to explain it, but since we put in 0.42 and 52 was a reasonable, or it wasn't an uncommon result, yeah, it wasn't significantly high or significantly low, then it's likely that this [0.42] could be the actual proportion. So, it should be included in the range.</p> <p>Lyla: We started by drawing a sample of undergraduates without knowing the true proportion of students that are intended business majors, which my value was 0.46. And then I just made a prediction what the true population may look like based on this sample. I didn't stray much from the 0.46, I did 0.5 and then made a distribution based on this 0.5 and saw where the probability of drawing 46 students in a sample of 100 from this larger population fell on that success rate. And then based on that, tried to find values where drawing 46 was the very furthest extremity and then the lowest extremity to try to find cases in which it seems nearly impossible to get 0.46, both it being the smallest number and the largest number. And from that, knowing the proportions in which it was basically impossible to get the number 46, we know where it is possible to have 46.</p>
Deductive	<p>Corrina: So, when looking at the dot plot, I would put in our prediction of maybe it's 50% and then look for that 51 on the dot plot. And then just obviously we looked at the other numbers, but if you were really looking for, is it this or not, you could just kind of look at 51 and see how likely that happened. So, normally I would just kind of go straight for that 51 and see on this one it's this teeny tiny dot, but on one of the other ones it was the most probable thing on there...Just kind of where 51 was sitting on the scale. The number that we put in here (points to the probability of success on the applet) kind of is usually towards the middle, just kind of how close it's getting to this sample [outcome of 51] was kind of deciding whether or not it ends up in the scale.</p> <p>Mindy: So, this [first] one was just kind of a random prediction, because it's super close to 49. And then we plugged in different ranges, and I determined whether or not 49 would fall on the graph. And because of that you can kind of determine whether or not you think 20 or 30 would actually be in the range...When we saw 20, 49 wasn't super likely to show up.</p>

unknown parameter based on the probability of one sample outcome (Component 2) indicated a sophisticated understanding of sampling distributions. Although these seven students engaged in the same activity—they assessed the plausibility of multiple values based on the probability of their one sample—not all showed evidence of coordinating two levels of distribution.

Drawing Inferences About a Population Without Using the Result of One Random Sample

Other than in their initial prediction for the unknown parameter, four students did not use the result of their one sample outcome to determine their range of plausible values for the true percentage of intended business majors at South University (see Table 10). Instead, Julie reasoned deductively based on the shape and highest peak of each sampling distribution (Component 4) she constructed using the web-based applet, and Jess reasoned inductively from the results of five random samples she drew from the box of beads (Component 3). The remaining two students—Becky and Waverly—did not show evidence of using any of the three broader forms of reasoning to determine their range of values for the true parameter at South University. Although all four students understood that random samples will produce outcomes that vary from the population parameter (from Interview 1 results), they did not show evidence of using this understanding in a new context—one in which the population parameter was unknown—to determine the likelihood of their sample outcome under multiple hypothesized values for the percentage of undergraduates at South University who are intended business majors. Furthermore, none of these four students showed evidence of relating one or more of the sampling distributions they constructed using the applet to a hypothetical population (Component 1) or assessing the plausibility of one or more values for the unknown population parameter based on the likelihood of their sample outcome (Component 2). I hypothesize that these four students viewed each distribution they constructed using the applet as a collection of

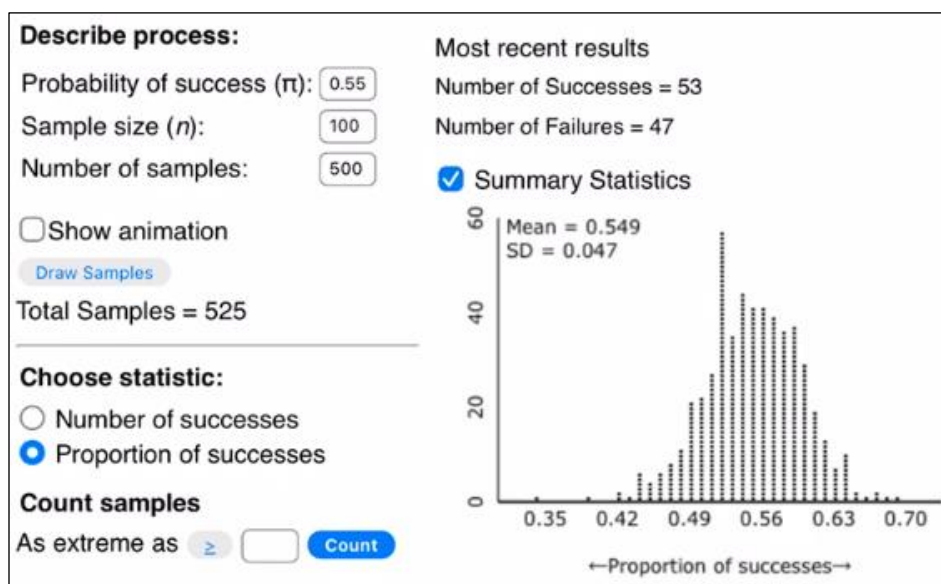
sample outcomes from the actual South University, rather than from a hypothetical population—or even from a hypothetical version of South University—whose parameter was equal to the probability of success they input. This is not unreasonable because, in Interview 1, the sampling distribution they constructed using the applet did represent a collection of sample outcomes from North University. Recall, a critical understanding in making inferences from sample data to a larger population is the idea that because sample outcomes vary from the population parameter, the one observed sample outcome is simply one case of many possible outcomes that could have been produced by a population with a number of different parameter values. These four students did not show evidence of this understanding. Although all four students identified the probability of success they entered in the applet as the “true proportion,” it is possible they did not understand that the applet was designed to draw samples from a population whose parameter is equal to the probability of success. Alternatively, it is possible that their understanding of sampling distribution did not enable them to use the applet in productive ways to examine possible values for the unknown population parameter.

Using Visual Characteristics of a Graph: Julie’s Deductive Reasoning. Julie was the only participant who determined her range of plausible values for the unknown parameter at South University using visual characteristics of the dot plots she constructed using the applet. Although Julie used her sample outcome in her initial prediction, she never used the likelihood of her sample outcome to assess the plausibility of particular values. Instead, Julie constructed multiple sampling distributions, identified the “highest peak,” and included her predicted values in her range if the “highest peak” was close to her predicted value. Like all other students, Julie chose an initial prediction for the unknown parameter that was close to her sample outcome; her one random sample resulted in 60% and she made an initial prediction of 55% for the unknown

parameter at South University. To test her initial prediction using the applet, Julie input 0.55 as the probability of success and drew 525 samples. The result of her simulation is shown in Figure 20. After asking Julie if any of the samples she drew resulted in samples like the one she drew from the box of beads, she said, “I definitely had at least five were at 0.6.” When I asked her what information that gave her about her initial prediction of 55%, she explained that her initial prediction of 55% was “more accurate” than her sample outcome because there were “more near 0.55 than 0.6.” Julie reasoned that, because there were more sample outcomes that resulted in 0.55 than 0.6, her initial prediction of 0.55 was a better estimate for the unknown parameter than her sample result of 0.6. This is evidence that she may have viewed the sampling distribution as a collection of sample outcomes from the actual South University. As I tried to understand Julie’s thinking, I questioned her about what the 0.55 that she input in the applet meant. She explained, “So, essentially if the university had reported the statistic and how many they believed were intended business majors, that's what the 0.55 is. If they had reported it, but right now it's

Figure 20

The Result of Julie’s Simulation When Testing 55%



what my prediction was.” I continued to question Julie to understand her thinking; I asked her about outcomes she would find surprising:

Claire: If this (points to the applet where 0.55 is entered as the probability of success) were really the true proportion of undergraduates at South University that were intended business majors, what would you be surprised about in a sample of 100?

Julie: I think I would be surprised at anything (looks at the dot plot) under 45% and anything above 63, 64%...just based on the way the samples are distributed on the dot plot.

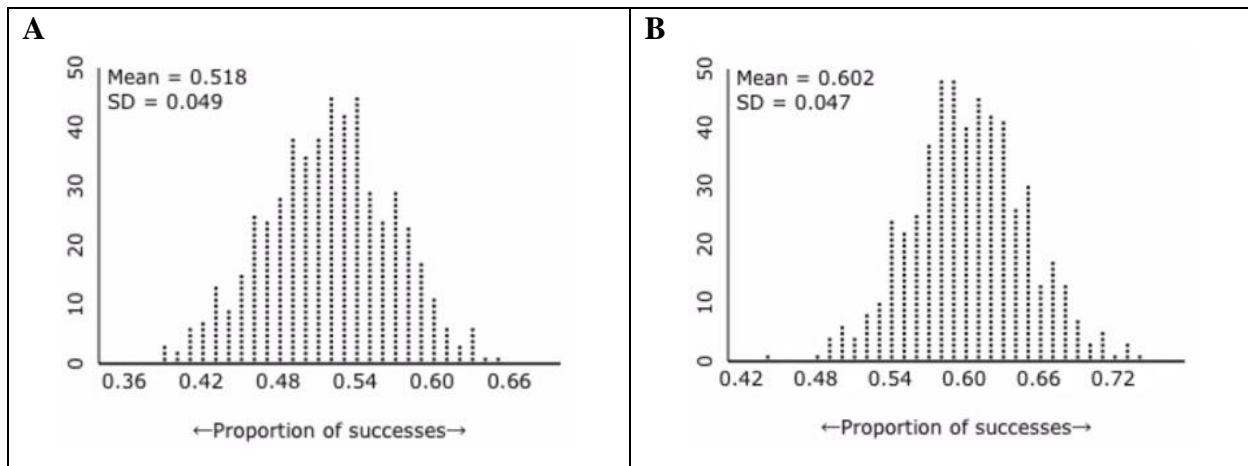
During the interview, I interpreted Julie’s description of the meaning of the 0.55 and her identification of surprising outcomes to mean she understood that she used the applet to draw 525 samples from South University, *if* South University’s true percentage of intended business majors was 55%. However, only through retrospective analysis do I now have a different understanding of Julie’s thinking; I hypothesize that she viewed each sampling distribution she constructed as a collection of sample outcomes from the actual South University. This would explain why Julie based some of her subsequent predictions on the mode, or what she called the “highest peak,” of one or more distributions she constructed with the applet. For example, after Julie tested her initial prediction of 55%, she chose to test 52%, which was the mode of the sampling distribution shown in Figure 20. Furthermore, she used the “highest peak” to help her determine whether to include her predicted value in her range of values for the unknown parameter. Rather than using the likelihood of her sample outcome, Julie constructed multiple sampling distributions, identified the “highest peak,” and included her predicted values in her range if the “highest peak” was close to her predicted value. From Interview 1, Julie understood that repeated random samples will produce outcomes that are centered at the true population

parameter. Thus, if Julie viewed each sampling distribution as a collection of outcomes from the actual South University, it makes sense that she identified the “highest peak” to estimate the true population parameter.

Julie tested multiple predictions, such as 52%, 60%, and 47%, among others. Each time Julie input a new prediction in the applet and drew 500 or more samples, she identified the mode(s) of the distribution and used this information to make future predictions or to determine whether she should include her predicted value in her interval estimate for the unknown parameter. For example, when Julie tested her prediction of 60%, she identified the modes in the distribution she constructed using the applet (see Figure 21). She said, “So, you have kind of two peaks. You've got a good bit at 58, 59, and then you've got some at 61 and 62, but you still have a significant amount at 60.” Although Julie included 60% in her range, she explained that “based on the dot plots and the highest peaks,” she was more confident in some of her previous predictions, such as 52%. She explained, “When I had it at those numbers, I had more centralized, I had more results, more of the samples that we did at that number.” Figure 21 displays the result of Julie’s simulation when she tested 52% and 60%. Julie attended to the highest peaks in each distribution; in the distribution for her prediction of 52% (Figure 21a), the two modes are 0.52 and 0.54; in the distribution for her prediction of 60% (Figure 21b), the two modes are 0.58 and 0.59. Julie explained that she was more confident that the true proportion for South University could be 52% rather than 60% because the mode of the distribution when she tested 52% was equal to 0.52; this was not the case in the distribution resulting for her testing 60%. Although she was less confident in her prediction of 60%, she still included it in her range of plausible values for the unknown population parameter.

Figure 21

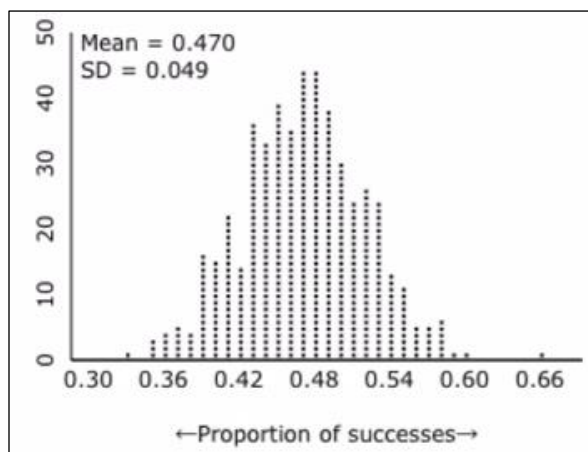
The Results of Julie's Simulations When Testing 52% and 60%



Julie used similar reasoning to include 47% in her range of values. She identified the modes when she said, “47 and 48 were tied for the highest” (see Figure 22). When she tested her prediction of 47%, I asked Julie if she would be surprised to get a sample outcome of 60, like the one she drew. She explained, “I would be surprised...because out of the 500 samples that we took, only one was at 60. So, I would be surprised.” When I asked her if this changed her opinion about 47%, she said she still liked 47% and included it in her range of values.

Figure 22

The Result of Julie's Simulation When Testing 47%



This is further evidence that Julie did not use the result of her sample outcome to determine plausible values for the true parameter at South University. Julie ultimately provided a range of 47% to 63% and summarized her process and justified her confidence by explaining, “Every time we do in the range of 47 to 63, it starts to centralize around that number. So that's why I am pretty confident it would be in that range.”

Julie’s repeated reasoning and summary suggest that she deduced her range of values from a rule involving the mode, or “highest peak,” of the distribution: If the mode is approximately equal to the probability of success, p , then p is a good estimate for the true parameter (rule). When testing her prediction of 47%, she identified 0.47 as being “tied for the highest” (case), so she deduced that 47% should be included in her range (result). Based on her understanding of how samples vary from the population parameter, if Julie viewed each sampling distribution as a collection of sample outcomes from South University, she would identify the center of the distribution as the parameter. Throughout her predicting and testing, Julie identified the outcome with the “highest peak” as a plausible value for the true parameter, supporting my hypothesis that she viewed each sampling distribution she constructed as a collection of sample outcomes from the actual South University, rather than from a hypothetical population whose parameter was equal to the probability of success she input.

Using the Result of Multiple Samples: Jess’s Inductive Reasoning. Jess was the only participant who determined her range of plausible values for the unknown parameter at South University using the results of multiple samples she drew from the box of beads. For her initial prediction, Jess chose 50% and said that she could have chosen 60% based on her sample of 59%, but explained, “I believe having 60% business majors just seemed like way too many” and thought “it was more of half and half.” Like Julie, Jess identified the mode of each distribution

and used this information to make future predictions. However, Jess was perturbed when she noticed that whenever she changed the probability of success to a different value, the distribution tended to shift and center around that value. She said,

I think in my head, the probability of success, it's like whatever number I'm going to change it to, it's going to kind of situate it around that. So, if I put it to 0.7, it's going to shift it to 0.7...Because I'm telling it there's a 70% chance of success, therefore it's more likely to be situated around 70...So, I guess thinking about that now, I feel like maybe there is a way using this graph to determine the proper prediction, but I wouldn't know how based on just one thing (points to her sample) because I'm telling it this is the prediction or this is how many business majors there are at the school.

From this excerpt, Jess understood that the probability of success affected where the distribution was centered. However, it is not clear if Jess understood that the probability of success represented the parameter for the population that produced the collection of sample outcomes. At this point in Task D, Jess explained that she was not confident in any value she tested using the applet and said, “I think for me to give a range and feel more confident, I would want to scoop out a couple more and see truly how many I get each time.” So, I allowed her to do so. Jess drew five random samples from the box of beads (cases), observed that her sample outcomes were in the “high fifties and low sixties” (results), so she induced a range of 57 to 62 for the true parameter of South University (rule).

Jess was the only student to draw multiple samples from the box of beads and use these sample outcomes to induce a range of values she thought was plausible for the unknown population parameter. Reasoning inductively in this way is productive, but unrealistic. In other words, Jess could easily draw multiple samples and estimate where they clustered because the

samples she drew came from a box of beads representing South University undergraduates. However, it is not feasible or cost-effective to draw multiple samples from a population of interest to estimate the unknown parameter.

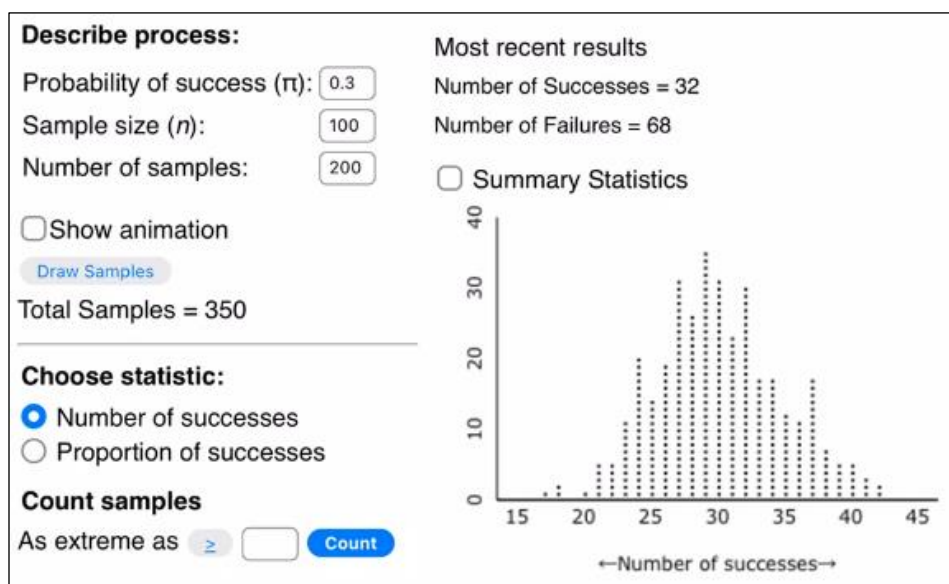
Other Ways of Reasoning. Neither Becky nor Waverly showed evidence of reasoning abductively, inductively, or deductively to determine their range of values for the unknown parameter. Furthermore, their reasons for choosing values to test were not always clear, making it difficult for me to articulate what they were thinking and what they understood. Both Becky and Waverly considered their sample outcome when using the applet to test multiple values for the unknown population parameter for South University, but their assessment of the plausibility of each value they tested was inconsistent. In other words, even if they identified their sample outcome as unlikely or an outlier under a hypothetical parameter, they still included that value in their range. On the other hand, if their sample outcome was likely to occur under a hypothetical parameter, they did not always include that value in their range. I provide an example of this inconsistency I identified in each student's reasoning below.

At the beginning of Task D, Becky drew a sample that produced an outcome of 57 and explained that she did not think the unknown parameter for South University was the same as North University's 20% because her sample outcome was much higher than 20%. Instead, she thought that the true parameter was closer to her outcome of 57 and gave an initial prediction of 55%. When she tested her prediction of 55% using the applet, she did not use the result of her simulation nor consider the likelihood of her sample outcome. Because Becky had already mentioned she did not think the parameter could be 20%, I wanted to direct her attention to a collection of sample outcomes in which her sample was extremely unusual. Thus, I asked Becky if she thought the true percentage of business majors at South University could be 30%.

Although Becky did consider her sample outcome of 57, her assessment of 30% as a plausible value did not align with my expectation. After inputting a probability of success of 0.30 in the applet, she drew 100 samples and noted, “I didn’t get [my] sample even in 100 times. Even if I did it more, I doubt it would be more than 45.” She continued to draw more samples until she had a collection of 350 sample outcomes and said, “Yeah, it’s still not past 45.” The result of her simulation is shown in Figure 23. Becky noted that the sample outcomes did not exceed 45 and that her sample outcome never occurred. At this point, I expected that she would rule out 30% as a plausible value for the unknown parameter because her sample outcome of 57 was nearly impossible to obtain under this hypothetical parameter value. However, when I asked her how she felt about 30% as a prediction for the true proportion after seeing the results of her simulation, she said she thought the unknown parameter “could be” 30% and included this value in her range of plausible values for the unknown parameter. Like in this case, there were other times throughout Task D that Becky identified her sample outcome of 57 as unlikely or an outlier

Figure 23

The Result of Becky’s Simulation When Testing 30%

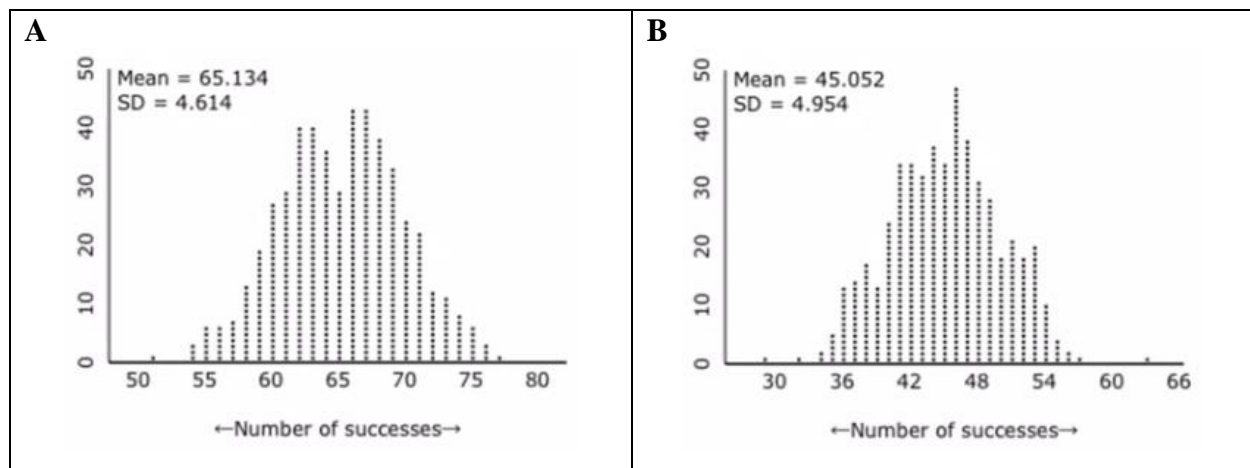


under additional hypothetical parameter values, such as 0.33 and 0.40, yet she still included those in her range for the unknown parameter; Becky's final range was 30% to 60%. These examples show that Becky did not use the likelihood of her sample outcome to determine her range of values for the unknown parameter at South University.

Waverly's inconsistent reasoning also made it difficult to understand what she was thinking and how she determined her final range of values. Although she mentioned the likelihood of her sample outcome under several values she input for the probability of success, she did not include or exclude her predictions in her range consistently based on the likelihood of her outcome. For example, when testing her prediction of 65%, Waverly input 0.65 as the probability of success, drew 500 samples, and noted that there were "only six" outcomes that resulted in her sample outcome of 56. Figure 24a shows the result of her simulation. When I asked Waverly what it meant that there were only six outcomes like hers, she said, "That it's possible, but not very common. But it's also not a huge outlier." Waverly tested 45% by inputting 0.45 as the probability of success and drawing 500 samples. In talking about the likelihood of her sample outcome of 56, she said, "It's not super likely, but it's possible. So, I can't say with any confidence that it's not right, because it's possible." Figure 24b shows the result of her simulation. Based on the results of Waverly's simulations shown in Figure 24, I would expect her to make the same decision about including or excluding 45% and 65% in her final range of values, had she been using the likelihood of her sample outcome to determine her range. However, Waverly's final range was 45% to 60%. Although Waverly showed some evidence of using the likelihood of her sample outcome to assess the plausibility of some values, she did so inconsistently. Furthermore, sometimes she attended to the shape of the distribution rather than the likelihood of her sample outcome. For example, she excluded both 30% and 80%

Figure 24

The Results of Waverly's Simulations When Testing 65% and 45%



as possible values for the unknown parameter based on the shape of the distribution; she ruled out 30% because the collection of outcomes was not “the right distribution shape...the bell-shaped curve” and she ruled out 80% because the distribution was “skewed.”

Like Julie and Jess, it is possible that Becky and Waverly viewed each sampling distribution as a collection of outcomes from the actual South University. However, the inconsistencies in both Becky’s and Waverly’s reasoning made it difficult to understand how they determined their range of values for the unknown population parameter. Although both Becky and Waverly noted that the probability of success represented the “true” proportion or percentage, they did not explicitly relate the sampling distributions they constructed using the applet to a hypothetical population that could have produced the collection of outcomes. Furthermore, neither Becky nor Waverly used the likelihood of their sample outcome consistently to determine their range of values for the unknown parameter at South University.

Summary

Looking across all 11 students, the three students who reasoned abductively showed a flexible understanding of sampling distribution because they were able to coordinate population

distribution and sampling distribution across two different contexts—one in which the parameter was known (Interview 1) and one in which the parameter was unknown (Interview 2). In Interview 1, many students showed evidence of understanding that the sampling distribution they constructed when $p = 0.2$ represented a collection of possible outcomes for a sample drawn from North University. However, when the population parameter was unknown in Interview 2, most of these students did not show an understanding that the distribution they constructed using the applet with probability of success, p , represented a collection of possible sample outcomes for a sample drawn from some hypothetical population with a parameter equal to p . I hypothesize that, in Interview 1, the connection between the sampling distribution they constructed with the applet and the population distribution for North University was more obvious to students because (1) they were able to see the population that represented North University (i.e., the box of beads) from which they drew samples, and (2) the sampling distribution they constructed with the applet looked similar to the those they constructed with five and 20 samples from the box of beads in that their sample outcomes tended to center around 20%, North University's true parameter. It is not unreasonable for students to think that a sampling distribution they constructed in Interview 2 represented a collection of sample outcomes from South University because this is what it represented in Interview 1, but for North University. However, a critical understanding in making inferences from sample data is the idea that because sample outcomes vary from the population parameter, the one observed sample outcome is simply one case of many possible outcomes that could have been produced by a population with a number of different parameter values.

CHAPTER 7

DISCUSSION

The purpose of this study was to examine the forms of inferential reasoning—deduction, induction, and abduction—novice statistics students employ when reasoning about and with sampling distributions. Furthermore, I investigated what these forms of reasoning revealed about their understanding of sampling distributions. The research questions that guided this study are as follows:

1. When given a population with a known parameter, what forms of reasoning do novice statistics students employ when determining what sample outcomes they expect and what sample outcomes they find surprising, and what do these forms of reasoning reveal about their reasoning *about* sampling distributions?
2. When given a population with an unknown parameter, what forms of reasoning do novice statistics students employ when making inferences from one sample outcome to the population from which it was drawn, and what do these forms of reasoning reveal about their reasoning *with* sampling distributions?

In this chapter, I first summarize and discuss the findings I presented in the previous two chapters, connecting the findings to my research questions and relevant literature. Next, I discuss the limitations of the study. Lastly, I discuss the implications of the findings and address what questions still remain about students' reasoning about and with sampling distributions and provide suggestions for directions of future research.

Reasoning About Sampling Distributions (Research Question 1)

When given a population with a known parameter, I found that the novice statistics students in my study employed both deductive and inductive reasoning when determining what sample outcomes they expect and what sample outcomes they find surprising. However, inductive reasoning was particularly useful in supporting students in developing a better understanding of sampling variability. Furthermore, students reasoned inductively only after having drawn physical samples from the box of beads and virtual samples using the web-based applet. They observed patterns in the collection of outcomes and then generalized these patterns to future outcomes. Prior to drawing samples, students tended to reason deductively based on informal facts or rules, likely recalled from their recently completed introductory statistics course. Unsurprisingly, most students admitted not remembering how to carry out formal calculations based on statistical formulas. However, reasoning inductively supported students in making more appropriate estimates for sampling variability.

To support students in attending to sampling variability, several researchers reported on their use of physical samples (e.g., Kelly & Watson, 2002; Shaughnessy et al., 2004; Torok & Watson, 2000). These studies investigated students' reasoning in repeated sampling situations, and researchers engaged students in making initial predictions, physically drawing samples, and then comparing initial predictions to actual sample outcomes. Although Kelly and Watson (2002) found that many students were hesitant to change their predictions after physically drawing samples, Shaughnessy et al. (1999) found evidence that physically taking samples increased the likelihood that a student would provide reasonable predictions for sample outcomes. My results align with and build on the findings of Shaughnessy et al. (1999). The students in my study tended to provide more reasonable estimates for 20 sample outcomes after

having engaged in the process of predicting the outcome of five random samples, drawing five random samples from the box of beads, and then comparing their predicted and actual sample outcomes. Not only did the students in my study tend to provide more reasonable predictions for sample outcomes after physically drawing samples, they also provided more reasonable estimates for sampling variability after having drawn both physical and virtual samples.

Prior to drawing samples, the students in my study provided a range of values they would expect to get in one random sample drawn from North University, a population with a known parameter. In doing this, they estimated how far they expected the sample outcome to vary from the population parameter. Recall, eight students provided an initial range of values that was centered at the population parameter with a reasonable margin while three provided unreasonably wide ranges that were not centered at the population parameter. Following the recommendations of Shaughnessy et al. (1999) and others, I intentionally engaged students in repeated opportunities to draw multiple samples to support them in attending to sampling variability. After drawing samples, students provided more reasonable ranges for the outcome of one random sample; thus, they provided better estimates for sampling variability. This was especially evident for the three students who provided unreasonably wide ranges prior to drawing samples. For example, Becky's initial range was 0% to 80% and her final range was 5% to 32%. For Becky and others, reasoning inductively by repeatedly observing patterns in collections of sample outcomes supported them in developing a better understanding of sampling variability.

The three tasks I designed to investigate the forms of reasoning students employed when reasoning *about* sampling distributions provided students with three opportunities to draw multiple samples, observe patterns in collections of sample outcomes, and generalize those

patterns to future sample outcomes. Engaging in these three repeated sampling simulations afforded students the opportunity to reason inductively to provide more reasonable ranges for the outcome of one random sample drawn from North University, a population with a known parameter.

Although reasoning inductively supported students in developing a better understanding of sampling variability, other aspects of students' reasoning indicated different understandings of sampling distribution. Recall, of the students who repeatedly reasoned inductively to update and refine their estimate for sampling variability, three students—Julie, Becky, and Corrina—showed evidence of a local understanding of data (Ben-Zvi, 2004) when they focused on individual observations rather than on characteristics of the entire distribution (e.g., shape, center, spread) when comparing their predicted and observed outcomes. Furthermore, these students viewed the empirical sampling distributions that were constructed from physically drawing random samples differently from those that were constructed from drawing random samples using the web-based simulation tool. They did not use the results of the 500 random samples they drew using the web-based applet to induce their final range of values. Instead, they used the results of their physically drawn samples. For these students, the extension from the physical simulation tool to the web-based tool was not obvious and, as a result, these students did not view the sampling distribution of 500 sample outcomes as the best approximation for the theoretical sampling distribution. Simulation is widely used in statistics education, with physical simulation tools often preceding web-based simulation tools. Although the connection between the models represented by these two tools is obvious to us as the teacher, my results show that not all students constructed the same meaning and understanding from the repeated sampling process using these simulation tools. More research is needed to examine how to support students in understanding these critical

connections and extensions and, as a result, developing a better understanding of complex and abstract statistical concepts, such as sampling distributions.

Reasoning With Sampling Distributions (Research Question 2)

When given a population with an unknown parameter, I found that novice statistics students in my study employed all three forms of reasoning when making inferences from one sample outcome to the population from which it was drawn. However, students who reasoned abductively showed evidence of having a more sophisticated and flexible understanding of sampling distributions than those who reasoned deductively or inductively. Students who reasoned abductively were able to reason about the multi-tiered repeated sampling process (Saldanha & Thompson, 2002) across two different contexts—one in which the population parameter was known (Interview 1) and one in which the population parameter was unknown (Interview 2). In contrast, students who reasoned deductively and inductively did not show evidence of coordinating multiple levels of distribution.

Consistent with the results of Saldanha and McAllister (2014), I found that a majority of the students in my study used their sample data to make informal inferences about the population. However, their ways of reasoning revealed differences in their understanding of sampling distributions. Of the students who used the likelihood of their sample outcome to determine a range of plausible values for the unknown population parameter, some reasoned abductively while others reasoned deductively. Students who reasoned abductively hypothesized multiple populations (and corresponding parameters) that could have produced their sample outcome. In contrast, students who reasoned deductively included the likelihood of their sample outcome in their rule for determining whether to include a particular prediction in their range of values for the unknown parameter. I identified a subtle yet critical distinction in students'

reasoning across these two groups; those who reasoned abductively coordinated two levels of distribution—population distribution and sampling distribution—and those who reasoned deductively did not show evidence of this coordination. This distinction was so subtle that I initially coded both groups of students as having reasoned abductively. In my initial interpretation of students' reasoning (later coded as deductive), I inferred they understood that they used the applet to draw samples from some hypothetical population each time they predicted a new value and constructed a sampling distribution with the applet. However, additional analysis and comparisons of students' reasoning clips and summaries (see Table 11) across the two groups highlighted explicit mentions of hypothetical populations in only one of the groups—the group who reasoned abductively.

Recall, I hypothesized that reasoning abductively would be a productive way to reason with sampling distributions to draw inferences from sample data to a population with an unknown parameter because abductive reasoning shares aspects of inferential reasoning. In statistical inference, we start with the outcome of a sample and, by considering the probability of the sample outcome (or one more extreme) under a hypothetical parameter (call this p), we conclude the sample outcome could be explained by the supposition that it was drawn from a population whose parameter is equal to p . Using Peirce's (1878) definitions of case, rule, and result, I argue this is an abductive inference, concluding a case from a rule and a result. In short, we start with a sample outcome (result) and, using probability, conclude that the sample could have been drawn from (is a case of) a population whose parameter is equal to p (general rule). A critical component in this abduction is that the conclusion of the inference is that the sample could have been drawn from a population whose parameter is equal to p . Based on the way I defined case, rule, and result in this abductive inference, this conclusion can only be made by

connecting a collection of sample outcomes to the population from which they were drawn. For the few students who reasoned abductively in my study, they explicitly connected each sampling distribution they constructed using the applet to a hypothetical population that could have produced that collection of sample outcomes. No other student made this connection explicit, nor did I interpret their reasoning as abductive. This result brings up an important question about the causal mechanism between coordinating multiple levels of distribution and abductive reasoning. Were students able to reason abductively *because* they understood the relationship between the hypothetical population and the sampling distribution, or did they understand the relationship between the hypothetical population and the sampling distribution *because* they reasoned abductively? My initial hypothesis was that these students were able to reason in this way because they understood the relationship between the hypothetical population and the sampling distribution. However, one student's final remarks at the end of Interview 2 gave me pause. Lyla said,

That was just really cool. I don't know. It was cool to go backwards from, you just know when you take stat, it's like, oh, we have this confidence that this number is going to fall in this population because of blah, blah, blah, blah. But working backwards shows the logic behind it even more. I thought it made sense to me before, it made perfect sense, our rationale for drawing these conclusions. But going backwards, it's another level of understanding.

Although I cannot make any claims about Lyla's understanding of drawing inferences from sample data prior to her involvement in this study, I do know that the ways in which confidence intervals are taught in the introductory statistics course she took do not engage students in repeatedly predicting and testing possible population parameters like the tasks that I designed for

this study did. Lyla's reflection on her own understanding is powerful. I can hypothesize (and hope) that Lyla's engagement in these tasks supported her in coordinating multiple levels of distribution, contributing to what she described as "another level of understanding." However, the question about the causal mechanism between this coordination and her abductive reasoning still remains.

Recall, not all students used the result of their sample outcome to determine a range of plausible values for the unknown population parameter. This was surprising because the result of one sample was the only information students had about the population to which they were making inferences, so I expected that students would use this information in their reasoning. I hypothesized that these students viewed each sampling distribution they constructed using the applet as a collection of outcomes from the actual South University, rather than from a hypothetical population (or a hypothetical South University) whose parameter was equal to the probability of success they input in the applet. It is not unreasonable for students to think this because, in Interview 1, the empirical sampling distributions they constructed from drawing samples from both the box of beads and using the applet represented a collection of possible sample outcomes from North University. From Interview 1, these students understood that random samples will produce outcomes that vary from and cluster at the population parameter but were unable to use this understanding in a context in which the population parameter was unknown (Interview 2). During the interview, I interpreted these students' identification of the probability of success they entered in the applet as the "true" proportion or percentage to mean that they understood that the applet was designed to draw samples from a population whose parameter was equal to the probability of success they entered. However, my retrospective analysis indicated otherwise. It is possible that these students' existing meanings did not enable

them to productively engage with the applet to further their understanding. Following the recommendations of several researchers (e.g., Garfield et al., 2012; Saldanha & Thompson, 2014), I incorporated simulation tools in my study to support students in visualizing the abstract and complex concept of sampling distribution. However, more research is needed to understand the meanings students construct when using these simulation tools.

Limitations

This study is limited by the small sample size, as I only characterized the reasoning of 11 students. It is possible that a larger group of students might show additional differences in their reasoning than the categories of reasoning I identified in this study. In addition, all 11 students had recently completed the same introductory statistics course at the same university. Although the instructors may have differed, the course was a coordinated course, with the expectation that all instructors use the same instructional materials and assessments. Furthermore, their introductory statistics course included a lab component in which students used both physical and virtual simulation tools to explore concepts like sampling distribution and statistical inference. Thus, it is possible that these students' prior experiences in their introductory statistics course could have affected their ways of reasoning about and with sampling distributions.

As part of my initial proposal for this study, I aimed to examine the forms of inferential reasoning students employed when distinguishing between three types of distribution—population distribution, sample distribution, and sampling distribution. However, I was unable to fully answer this question from the data that I collected. At the end of Interview 1, I asked the participants to draw a picture that represented the repeated sampling process that the applet modeled for the North University context (see APPENDIX C, prompt 23). I intended for these drawings to help me understand how students distinguished between the three distributions, but I

identified two reasons why I was unable to do this. First, most students simply reproduced the dot plot that was displayed on the applet, and their drawings did not provide insight into how they were thinking about the different distributions. I did not expect students to replicate what they produced using the web-based applet because when I piloted my interview protocols, the students I interviewed did not do this. It is possible that the students in my study tended to replicate what they saw on the applet because I asked them to draw their picture directly following their exploration with the simulation tool, and the dot plot displayed on the applet may have been the first thing that came to mind as a potential drawing to produce. Second, although I asked students to explain their drawings, I did not ask sufficient follow up questions due to the limited amount of time remaining in the interview. If I were to repeat this study, I would explicitly ask students to explain the relationship between the distribution of business majors at North University, the distribution of business majors in one random sample, and the distribution of the sample proportion. Although I was not able to fully capture how students distinguished between the different distributions from their drawings and explanations in Task C, students' reasoning in Interview 2 provided some insight into how they reasoned about the relationship between population distribution and sampling distribution, as I described in the previous chapter. Thus, I adjusted my research questions to better reflect the phenomena my data captured.

For students who used the likelihood of their sample outcome to determine a range of values for the unknown population parameter at South University in Interview 2, I identified a subtle and critical difference in their explanations—some students explicitly related the sampling distributions they constructed using the applet to the hypothetical distribution that produced the collection of outcomes. Based on my interpretation of case, rule, and result in this context, I identified these students as having reasoned abductively. In contrast, for students who did not

show evidence of coordinating multiple levels of distribution, I identified their reasoning as deductive. It is possible that the students who used the likelihood of their sample outcome to deduce their range of values for the unknown parameter did consider the relationship between the population distribution and the sampling distribution, but did not show explicit evidence of this with their gestures or verbal explanations. Thus, my interpretation of these students' reasoning may be limited by the fact that I characterized their reasoning based on a *lack* of evidence.

I designed my study to incorporate the use of simulation tools to support students in visualizing the abstract and complex repeated sampling process involved in constructing a sampling distribution. Whether it was the design of the applet, students' ways of reasoning, or a combination of both, a subset of my participants were unable to engage in ways that were productive for their understanding. Although all students had prior experience with one or more web-based simulation tools in their introductory statistics course, none of the students had experience with the applet I used in this study. The applet functioned similarly to the tools used in the introductory statistics course these students had recently taken, but it is possible that the students' unfamiliarity with the tool and the short time they had available to explore the tool prevented them from engaging with the applet productively. During Interview 2, all students referred to the probability of success they entered in the applet as a "true" proportion or percentage and, at the time of data collection, I took this to mean that the students understood that they were using the applet to draw samples from a hypothetical population with a parameter equal to the probability of success shown in the applet. However, the results of Interview 2 indicated that most students did not view the applet in this way. Had I known this during data collection, I would have adjusted how I interacted with the students as they engaged with the

applet. Namely, in addition to asking the students to explain how the applet represented the sampling context, I would ask them to describe the relationship between the probability of success they entered and the collection of sample outcomes they produced using the applet. Students' descriptions of this relationship could provide insight into how or if they coordinated multiple levels of distribution. It is also possible that other simulation tools may be more effective in supporting students in reasoning productively about repeated sampling processes. For example, a simulation tool that displays all three levels of distribution—population distribution, sample distribution, and sampling distribution—may be more effective in supporting students in visualizing the multi-tiered sampling process.

Implications for Research and Teaching

Previous studies addressing students' understanding of sampling distribution characterized students' conceptions of sample from a "part-whole" perspective and described hierarchical levels of reasoning about sampling variability (e.g., Kelly & Watson, 2002; Saldanha & Thompson, 2002; Shaughnessy et al., 2004). In my research, I borrowed ideas from both Conner et al. (2014) and Conner and Peters (2023) and took a different approach to investigate novice statistics' students reasoning about and with sampling distribution by adopting Peirce's (1878) existing definitions of deduction, induction, and abduction that he described by the triadic structure of case, rule, and result. This approach allowed me to understand how these three broad forms of inferential reasoning are useful in distinguishing between the critical aspects of statistical reasoning and mathematical reasoning. In particular, I defined statistical inference in terms of abductive reasoning, which allowed me to understand how students coordinated multiple levels of distribution to draw inferences from sample data to a larger population. This approach of using Peirce's (1878) forms of inferential reasoning could be

adopted by other researchers to investigate students' understanding of other important statistical concepts and statistical reasoning more broadly.

To motivate this study, I argued that the forms of reasoning that are expected and highlighted in statistics differ from those that are standard in mathematics. Mathematical reasoning is often deterministic, emphasizing proof and deduction. In contrast, statistical reasoning is about reasoning probabilistically, *not* deterministically. Although all three forms of reasoning—deduction, induction, and abduction—are seen in mathematics, the emphasis is often on moving from conjectures and hypotheses to deductive proof. In mathematics, abductive and inductive reasoning can be used as initial or intermediary steps in proving processes (Pedemonte & Reid, 2011; Reid, 2018), but the end goal is often deductive reasoning and the construction of a mathematical proof. However, in statistics, reasoning about data necessarily means the conclusions drawn are uncertain. Thus, in statistical reasoning, the end goal is not deduction, but rather induction or abduction. The results of this study highlight this important distinction and have important implications for the teaching and learning of statistics.

The Common Core State Standards in Mathematics (CCSS-M) address ideas of sampling variability in the statistics and probability standards in as early as the sixth grade (NGACBP & CCSSO, 2010). My findings indicate that inductive reasoning supported students in developing a better understanding of sampling variability. Thus, it is important to provide students with opportunities to reason in this way. Because inductive reasoning involves generalizing patterns from observations, teachers can support students in reasoning inductively by engaging them in tasks that involve collections of sample outcomes and intentionally asking students to describe the patterns that they notice. In earlier grades, teachers can provide students with opportunities to make predictions about sample outcomes, draw sample outcomes, then compare their

observations with their predictions to support students in attending to sampling variability. This process of predicting and comparing can be extended in later grades to include explorations with adjusting sample size. For example, designing tasks in which students predict and compare sample outcomes for multiple different sample sizes can support them in understanding the effect of sample size on sampling variability.

Building on ideas of sampling variability and the relationship between a sample and the population from which it was drawn, standards in statistics and probability at the high school level include making inferences from sample data to a population of interest. Students are expected to investigate and interpret variability in repeated sampling situations to draw conclusions about data (NGACBP & CCSSO, 2010). Furthermore, the focus of introductory statistics courses at both the undergraduate and graduate levels is on statistical inference, whether informal or formal. The results of this study showed that abductive reasoning was powerful in making inferences from sample data to a larger population. Providing students with opportunities to reason abductively can support them in understanding the underlying logic and processes in determining a range of plausible values for an unknown population parameter. For example, providing students with opportunities to continually hypothesize multiple populations from which a sample may have been drawn can promote abductive reasoning. Engaging with simulation tools can promote these critical forms of reasoning, but more research is needed to understand the meanings students construct when using these tools. I discuss this in more detail in the next section.

In addition to providing students with opportunities to engage in critical forms of reasoning, such as inductive and abductive reasoning, I believe teachers should facilitate explicit discussions with students about the uncertainty involved in reasoning about data. The majority of

students' experiences in their school mathematics courses involve deterministic thinking and deductive reasoning, so students may feel uncomfortable and uneasy about drawing uncertain conclusions from data. Thus, having explicit conversations about the different kinds of conclusions drawn in mathematics and statistics can support students in feeling more comfortable with the nondeterministic nature of statistical reasoning.

Directions for Future Research

In this study, I found that both inductive and abductive reasoning were critical in reasoning about and with sampling distributions. Inductive reasoning is commonly defined as generalizing patterns from observations, and describing patterns is an important aspect of making sense of and analyzing data. For example, identifying clusters in a set of data is important in describing measures of center, such as mean and mode. Abductive reasoning can be thought of as hypothesizing what rule might produce a particular observation, and I found that students who reasoned in this way had a sophisticated understanding of sampling distribution. These results raise an important question about the role of inductive and abductive reasoning in students' understanding of other important concepts in statistics. Future studies should investigate how these forms of reasoning that are critical in understanding sampling distributions can support students' statistical reasoning more broadly. In addition, it may be productive to explore how different kinds of abduction, such as Eco's (1983) undercoded and overcoded abduction and Magnani's (2001) creative and selective abduction, are useful in characterizing key aspects of statistical thinking.

One way to promote inductive and abductive reasoning is by engaging students in the use of simulation tools, like those I used in this study. Since the early 2000s, statistics educators and researchers have recommended the use of simulation tools to support students' understanding of

sampling distributions and statistical inference (e.g., Garfield & Ben-Zvi, 2008; Noll & Shaughnessy, 2012). However, few studies have directly addressed the effect of these simulation tools on students' understanding. Furthermore, although simulations can provide dynamic visualizations of abstract concepts such as sampling distribution and statistical inference, students may not necessarily understand the relationships between the patterns they see (Chance et al., 2004). In recent years, there has been a move toward using simulation-based curricula in introductory statistics courses (Loy, 2021). However, recent studies that have compared simulation-based and traditional curricula for introductory statistics courses have found that students who received the simulation-based curriculum performed the same as or marginally better on assessment items than students who received the traditional curriculum (e.g., Hildreth et al., 2018; Maurer & Lock, 2016).

As a former statistics teacher, I often engaged my students in using both physical and web-based simulation tools to model practical sampling scenarios. As teachers, we understand the connections between a sampling scenario and the simulation tools that model the scenario. I found that using simulation tools can promote critical forms of reasoning, such as inductive and abductive reasoning. However, I also found that students may view a sampling scenario and the simulation tool modeling that scenario differently, suggesting they may not understand the underlying mathematical models that these simulation tools approximate and visually represent. Moreover, although experts understand the extension from an empirical sampling distribution constructed using a physical simulation tool to one constructed using a web-based simulation tool, I found that novice statistics students may not. In particular, they may not see that the models created by drawing a large number of samples using a web-based simulation tool better approximate the theoretical sampling distribution compared to those created by using physical

simulation tools. Simulation tools can be powerful in making complex and abstract ideas more concrete for students, but more research is needed to examine how to support students in understanding these critical connections and extensions and, as a result, developing a better understanding of complex and abstract statistical concepts.

The results of my study generated additional questions about the use of simulation tools. For example, do students need to already understand the underlying mathematical models that simulation tools approximate to meaningfully engage with them, or can teachers use simulation tools to help students establish this understanding? Whether it was the design of the applet, students' extant ways of reasoning, or a combination of both, a subset of my participants were unable to engage in ways that were productive for their understanding. Thus, future studies should examine the meanings students construct as they engage with simulation tools, the prerequisite meanings students need for those tools to be productive in furthering their understanding, and how the design of these tools can support or hinder students' understanding.

With these new understandings, future research can investigate how teachers can implement simulation tools to support student learning at the classroom level. The use of simulation tools and the instructional design of lessons leveraging them must emerge from our understanding of students' reasoning and the ways in which their reasoning affords and constrains their tool use. Thus, a long-term goal of this work is to develop research-based resources and instructional practices that promote critical forms of reasoning, such as inductive and abductive reasoning, to support students in constructing important meanings and understandings of statistical concepts. Helping students develop strong statistical reasoning skills will prepare them to evaluate evidence and claims based on data, enabling them to become better educated, well-informed members of society.

REFERENCES

- Abrahamson, D. (2012). Rethinking intensive quantities via guided mediated abduction. *Journal of the Learning Sciences*, 21(4), 626-649. <https://doi.org/10.1080/10508406.2011.633838>
- Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal*, 3(2), 64-83. <https://doi.org/https://doi.org/10.52041/serj.v3i2.552>
- Bakker, A., & Gravemeijer, K. P. E. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 147-168). Springer Dordrecht. https://doi.org/https://doi.org/10.1007/1-4020-2278-6_7
- Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. A. (2020). *Pre-K-12 guidelines for assessment and instruction in statistics education II (GAISE II): A framework for statistics and data science education*. American Statistical Association. https://www.amstat.org/docs/default-source/amstat-documents/gaiseiiprek-12_full.pdf
- Bell, A. W. (1976). A study of pupils' proof-explanations in mathematical situations. *Educational Studies in Mathematics*, 7(1/2), 23-40. <http://www.jstor.org/stable/3481809>
- Ben-Zvi, D. (2004). Reasoning about data analysis. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 121-145). Springer Dordrecht. https://doi.org/10.1007/1-4020-2278-6_6
- Ben-Zvi, D., Bakker, A., & Makar, K. (2015). Learning to reason from samples. *Educational Studies in Mathematics*, 88(3), 291-303. <https://doi.org/10.1007/s10649-015-9593-3>

- Ben-Zvi, D., & Garfield, J. B. (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). Springer Dordrecht. https://doi.org/10.1007/1-4020-2278-6_1
- Chance, B., delMas, R., & Garfield, J. B. (2004). Reasoning about sampling distributions In D. Ben-Zvi & J. B. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295-323). Springer Dordrecht. https://doi.org/10.1007/1-4020-2278-6_13
- Clement, J. (2000). Analysis of clinical interviews: Foundations and model viability. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 547-589). Lawrence Erlbaum Associates.
- Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, 104(9), 801. <https://doi.org/10.2307/2975286>
- Cobb, P., & Steffe, L. P. (1983). The constructivist researcher as teacher and model builder. *Journal for Research in Mathematics Education*, 14(2), 83-94. <https://doi.org/10.5951/jresmetheduc.14.2.0083>
- Conner, A., & Peters, S. A. (2023). Distinctive aspects of reasoning in statistics and mathematics: Implications for classroom arguments. In G. F. Burrill, L. de Oliveria Souza, & E. Reston (Eds.), *Research on reasoning with data and statistical thinking: International perspectives* (pp. 259-278). Springer International Publishing. https://doi.org/10.1007/978-3-031-29459-4_20

- Conner, A., Singletary, L. M., Smith, R. C., Wagner, P. A., & Francisco, R. T. (2014). Identifying kinds of reasoning in collective argumentation. *Mathematical Thinking and Learning*, 16(3), 181-200. <https://doi.org/10.1080/10986065.2014.921131>
- de Villiers, M. (1990). The role and function of proof in mathematics. *Pythagoras*, 24, 17-24.
- delMas, R., Garfield, J., & Chance, B. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education*, 7(3). <https://doi.org/10.1080/10691898.1999.12131279>
- diSessa, A. A. (2007). An interactional analysis of clinical interviewing. *Cognition and Instruction*, 25(4), 523-565. <https://doi.org/10.1080/07370000701632413>
- Eco, U. (1983). Horns, hooves, insteps: Some hypotheses on three types of abduction. In U. Eco & T. Sebeok (Eds.), *The sign of three: Dupin, Holmes, Peirce* (pp. 198-220). Indiana University Press.
- Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, 5(1), 75-98. <https://doi.org/10.1177/0959354395051004>
- Findley, K., & Lyford, A. (2019). Investigating students' reasoning about sampling distributions through a resource perspective. *Statistics Education Research Journal*, 18(1), 26-45. <https://doi.org/10.52041/serj.v18i1.148>
- Franklin, C., Bargagliotti, A. E., Case, C. A., Kader, G. D., Scheaffer, R., & Spangler, D. (2015). *The statistical education of teachers*. American Statistical Association. <https://www.amstat.org/asa/files/pdfs/EDU-SET.pdf>
- Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-*

K-12 curriculum framework American Statistical Association.

https://www.amstat.org/docs/default-source/amstat-documents/gaiseprek-12_full.pdf

GAISE College Report ASA Revision Committee. (2016). *Guidelines for assessment and instruction in statistics education (GAISE) college report 2016*. American Statistical Association. https://www.amstat.org/docs/default-source/amstat-documents/gaisecollege_full.pdf

Galotti, K. M. (1989). Approaches to studying formal and everyday reasoning. *Psychological Bulletin*, 105(3), 331-351. <https://doi.org/10.1037/0033-2909.105.3.331>

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education*, 10(3). <https://doi.org/10.1080/10691898.2002.11910676>

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review*, 75(3), 372-396. <https://doi.org/10.1111/j.1751-5823.2007.00029.x>

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM*(7), 883-898. <https://doi.org/10.1007/s11858-012-0447-5>

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

Garfield, J. B., Le, L., Zieffler, A., & Ben-Zvi, D. (2015). Developing students' reasoning about samples and sampling variability as a path to expert statistical thinking. *Educational Studies in Mathematics*, 88(3), 327-342. <https://doi.org/10.1007/s10649-014-9541-7>

- Ginsburg, H. P. (1981). The clinical interview in psychological research on mathematical thinking: Aims, rationales, techniques. *For the Learning of Mathematics*, 1(3), 4-11.
<https://www.jstor.org/stable/40247721>
- Ginsburg, H. P. (1997). *Entering the child's mind: The clinical interview in psychological research and practice*. Cambridge University Press.
- Goldin, G. A. (2000). A scientific perspective on structured, task-based interviews in mathematics education research. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 517-545). Lawrence Erlbaum Associates.
- Groth, R. E. (2015). Working at the boundaries of mathematics education and statistics education communities of practice. *Journal for Research in Mathematics Education*, 46(1), 4-16.
<https://doi.org/10.5951/jresmetheduc.46.1.0004>
- Harel, G. (2014). Deductive reasoning in mathematics education. In S. Lerman (Ed.), *Encyclopedia of Mathematics Education* (pp. 143-147): Springer.
- Harel, G., & Weber, K. (2020). Deductive reasoning in mathematics education. In S. Lerman (Ed.), *Encyclopedia of Mathematics Education* (pp. 183-190). Springer International Publishing. https://doi.org/10.1007/978-3-319-77487-9_43-6
- Harradine, A., Batanero, C., & Rossman, A. (2011). Students and teachers' knowledge of sampling and inference. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics-challenges for teaching and teacher education: A joint ICMI/IASE study: The 18th ICMI study* (pp. 235-246). Springer Netherlands.
https://doi.org/10.1007/978-94-007-1131-0_24

- Heid, M. K., Perkinson, D., Peters, S. A., & Fratto, C. L. (2005). Making and managing distinctions: The case of sampling distributions. In G. M. Lloyd, M. Wilson, J. L. M. Wilkins, & S. L. Behm (Eds.), *Proceedings of the 27th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. PME-NA.
- Hildreth, L. A., Robison-Cox, J., & Schmidt, J. (2018). Comparing student success and understanding in introductory statistics under consensus and simulation-based curricula. *Statistics Education Research Journal*, 17(1), 103-120.
<https://doi.org/10.52041/serj.v17i1.178>
- Hodgson, T., & Burke, M. (2000). On simulation and the teaching of statistics. *Teaching Statistics*, 22(3), 91-96. <https://doi.org/10.1111/1467-9639.00033>
- Hunting, R. P. (1997). Clinical interview methods in mathematics education research and practice. *The Journal of Mathematical Behavior*, 16(2), 145-265.
- Jeannotte, D., & Kieran, C. (2017). A conceptual model of mathematical reasoning for school mathematics. *Educational Studies in Mathematics*, 96(1), 1-16.
<https://doi.org/10.1007/s10649-017-9761-8>
- Kadijevich, D., Kokol-Voljc, V., & Lavicza, Z. (2008). Towards a suitable designed instruction on statistical reasoning: Understanding sampling distribution with technology. In C. Batanero, G. Burrill, C. Reading, & A. Rossman (Eds.), *Proceedings of the ICMI Study 18 and 2008 IASE Round Table Conference*.
- Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition*, 11(2), 143-157.
[https://doi.org/10.1016/0010-0277\(82\)90023-3](https://doi.org/10.1016/0010-0277(82)90023-3)

- Kelly, B. A., & Watson, J. M. (2002). Variation in a chance sampling setting: The lollies task. In B. Barton, K. C. Irvin, M. Pfannkuck, & M. J. Thomas (Eds.), *Proceedings of the 25th annual conference of the Mathematics Education Research Group of Australasia: Mathematics education in the South Pacific, Auckland* (Vol. 2, pp. 366-373). MERGA.
- Kollosche, D. (2021). Styles of reasoning for mathematics education. *Educational Studies in Mathematics*, 107(3), 471-486. <https://doi.org/10.1007/s10649-021-10046-z>
- Konold, C., Higgins, T., Russell, S., & Khalil, K. (2015). Data seen through different lenses. *Educational Studies in Mathematics*, 88(3), 305-325. <https://doi.org/10.1007/s10649-013-9529-8>
- Konold, C., & Pollatsek, A. (2004). Conceptualizing an average as a stable feature of a noisy process. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 169-200). Springer Dordrecht. https://doi.org/10.1007/1-4020-2278-6_8
- Krueger, J. I., & Heck, P. R. (2017). The heuristic value of p in inductive statistical inference. *Frontiers in Psychology*, 8. <https://doi.org/10.3389/fpsyg.2017.00908>
- Lane, D. M. (2015). Simulations of the sampling distribution of the mean do not necessarily mislead and can facilitate learning. *Journal of Statistics Education*, 23(2). <https://doi.org/10.1080/10691898.2015.11889738>
- Lipson, K. (2002). The role of computer based technology in developing understanding of the concept of sampling distribution. In *Proceedings of the Sixth International Conference on Teaching Statistics*.

- Lipson, K. (2003). The role of the sampling distribution in understanding statistical inference. *Mathematics Education Research Journal*, 15(3), 270-287.
<https://doi.org/10.1007/BF03217383>
- Lithner, J. (2000). Mathematical reasoning in task solving. *Educational Studies in Mathematics*, 41(2), 165-190. <http://www.jstor.org/stable/3483188>
- Loy, A. (2021). Bringing visual inference to the classroom. *Journal of Statistics and Data Science Education*, 29(2), 171-182. <https://doi.org/10.1080/26939169.2021.1920866>
- Magnani, L. (2001). *Abduction, reason, and science: Processes of discovery and explanation*. Kluwer Academic/Plenum Publishers.
- Magnani, L. (2009). *Abductive cognition: The epistemological and eco-cognitive dimensions of hypothetical reasoning* (Vol. 3). Springer-Verlag. <https://doi.org/10.1007/978-3-642-03631-6>
- Maurer, K., & Lock, D. (2016). Comparison of learning outcomes for simulation-based and traditional inference curricula in a designed educational experiment. *Technology Innovations in Statistics Education*, 9(1). <https://doi.org/10.5070/T591026161>
- National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Author.
- National Council of Teachers of Mathematics. (2009). *Reasoning and sense making*. Author.
- National Governors Association Center for Best Practices and Council of Chief State School Officers. (2010). *Common Core State Standards (Mathematics)*. Author.
- Noll, J., & Hancock, S. (2015). Proper and paradigmatic metonymy as a lens for characterizing student conceptions of distributions and sampling. *Educational Studies in Mathematics*, 88(3), 361-383. <https://doi.org/10.1007/s10649-014-9547-1>

- Noll, J., Redmond, S., & Dolor, J. (2010). From beans to polls: Does understanding of statistical inference within a known population context transfer to an unknown population context? In *Proceedings of the 13th annual conference on research in undergraduate mathematics education*.
- Noll, J., & Shaughnessy, M. (2012). Aspects of students' reasoning about variation in empirical sampling distributions. *Journal for Research in Mathematics Education*, 43(5), 509-556. <https://doi.org/10.5951/jresmetheduc.43.5.0509>
- Noll, J., Shaughnessy, M., & Ciancetta, M. (2010). Students' statistical reasoning about distribution across grade levels: A look from middle school through graduate school. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the 8th International Conference on Teaching Statistics*. International Statistics Institute.
- Oaksford, M., & Chater, N. (2020). New paradigms in the psychology of reasoning. *Annual Review of Psychology*, 71, 305-330. <https://doi.org/10.1146/annurev-psych-010419-051132>
- Office of Instruction. (2022). *UGA Bulletin*. Retrieved March 26, 2023 from <http://bulletin.uga.edu>
- Pedemonte, B., & Reid, D. (2011). The role of abduction in proving processes. *Educational Studies in Mathematics*, 76(3), 281-303. <https://doi.org/10.1007/s10649-010-9275-0>
- Peirce, C. S. (1878). Illustrations of the Logic of Science VI: Deduction, Induction, and Hypothesis. *The Popular Science Monthly*, 13(August), 470-482.
- Peirce, C. S. (1960). *Collected papers*. Harvard University Press.

- Prodromou, T., & Pratt, D. (2006). The role of causality in the co-ordination of two perspectives on distribution within a virtual simulation. *Statistics Education Research Journal*, 5(2), 69-88. <https://doi.org/10.52041/serj.v5i2.501>
- Reading, C., & Shaughnessy, J. M. (2000). Student perceptions of variation in a sampling situation. In T. Nakahara & M. Koyama (Eds.), *Proceedings of the 24th conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 89-96). Hiroshima University.
- Reading, C., & Shaughnessy, J. M. (2004). Reasoning about variation. In J. Garfield & D. Ben-Zvi (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 201-226). Springer Dordrecht. https://doi.org/10.1007/1-4020-2278-6_9
- Reid, D. A. (2018). Abductive reasoning in mathematics education: Approaches to and theorisations of a complex idea. *Eurasia Journal of Mathematics, Science and Technology Education*, 14(9), em1584.
- Reid, D. A., & Knipping, C. (2010). *Proof in mathematics education: Research, learning and teaching*. Sense.
- Roche, D. (2022, December 2). Herschel Walker vs. Raphael Warnock polls predict Georgia runoff winner. *Newsweek*. <https://www.newsweek.com/herschel-walker-raphael-warnock-polls-predict-georgia-runoff-winner-1764136>
- Rossman, A., & Chance, B. (n.d.). *Rossman/Chance applet collection*. <http://www.rossmanchance.com/applets/>
- Rossman, A., Chance, B., & Medina, E. (2006). Some important comparisons between statistics and mathematics, and why teachers should care. In G. F. Burrill (Ed.), *Thinking and*

- reasoning with data and chance: 68th yearbook* (pp. 323-333). National Council of Teachers of Mathematics.
- Rubin, A. (2020). Learning to reason with data: How did we get here and what do we know? *Journal of the Learning Sciences*, 29(1), 154-164.
<https://doi.org/10.1080/10508406.2019.1705665>
- Rubin, A., Bruce, B., & Tenney, Y. (1991). Learning about sampling: Trouble at the core of statistics. In D. Vere-Jones (Ed.), *Proceedings of the 3rd International Conference on Teaching Statistics* (Vol. 1, pp. 314-319). International Statistical Institute.
- Saldanha, L. A. (2004). *"Is this sample unusual?": An investigation of students exploring connections between sampling distributions and statistical inference* [Unpublished doctoral dissertation]. Vanderbilt University.
- Saldanha, L. A., & McAllister, M. (2014). Using re-sampling and sampling variability in an applied context as a basis for making statistical inferences with confidence. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9)*. International Statistical Institute.
- Saldanha, L. A., & Thompson, P. W. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics*, 51(3), 257-270.
<https://doi.org/10.1023/A:1023692604014>
- Saldanha, L. A., & Thompson, P. W. (2007). Exploring connections between sampling distributions and statistical inference: An analysis of students' engagement and thinking in the context of instruction involving repeated sampling. *International Electronic Journal of Mathematics Education*, 2(3), 270-297. <https://doi.org/10.29333/iejme/213>

- Saldanha, L. A., & Thompson, P. W. (2014). Conceptual issues in understanding the inner logic of statistical inference: Insights from two teaching experiments. *The Journal of Mathematical Behavior*, 35, 1-30. <https://doi.org/10.1016/j.jmathb.2014.03.001>
- Scheaffer, R. L. (2003). Statistics and quantitative literacy. In B. L. Madison & L. A. Steen (Eds.), *Quantitative literacy: Why numeracy matters for schools and colleges* (pp. 145-152). National Council on Education and the Disciplines.
- Scheaffer, R. L. (2006). Statistics and mathematics: On making a happy marriage. In G. F. Burrill (Ed.), *Thinking and reasoning with data and chance: 68th yearbook* (pp. 309-321). National Council of Teachers of Mathematics.
- Schurz, G. (2017). Patterns of Abductive Inference. In L. Magnani & T. Bertolotti (Eds.), *Springer handbook of model-based science* (pp. 151-173). Springer International Publishing. https://doi.org/10.1007/978-3-319-30526-4_7
- Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester Jr. (Ed.), *Second handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (Vol. 2, pp. 958-1009). Information Age Publishing.
- Shaughnessy, J. M., Ciancetta, M., & Canada, D. (2004). Types of student reasoning on sampling tasks. In *Proceedings of the 28th conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 177-184).
- Shaughnessy, J. M., Watson, J. M., Moritz, J. B., & Reading, C. (1999). School mathematics students' acknowledgement of statistical variation. In C. Maher (Ed.), *There's more to life than centers. Paper presented at the research pre-session of the 77th annual meeting of the National Council of Teachers of Mathematics*. (Vol. 1).

- Shaughnessy, M., Chance, B. L., & Kranendonk, H. (2009). *Reasoning and sense making: Statistics and probability*. National Council of Teachers of Mathematics.
- Simon, M. A. (2019). Analyzing qualitative data in mathematics education. In K. R. Leatham (Ed.), *Designing, conducting, and publishing quality research in mathematics education* (pp. 111-122). Springer International Publishing. https://doi.org/10.1007/978-3-030-23505-5_8
- Sotos, A. E. C., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*, 2(2), 98-113. <https://doi.org/10.1016/j.edurev.2007.04.001>
- Starnes, D. S., Tabor, J., Yates, D. S., & Moore, D. S. (2014). More about regression. In *The practice of statistics* (5 ed., pp. 737-800). W. H. Freeman and Company.
- Steffe, L. P., & Thompson, P. W. (2000). Teaching experiment methodology: Underlying principles and essential elements. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 267-306). Lawrence Erlbaum Associates.
- Stylianides, G. J., & Stylianides, A. J. (2008). Proof in school mathematics: Insights from psychological research into students' ability for deductive reasoning. *Mathematical Thinking and Learning*, 10(2), 103-133. <https://doi.org/10.1080/10986060701854425>
- Swanson, D., Schwartz, R., Ginsburg, H. P., & Kossan, N. (1981). The clinical interview: Validity, reliability and diagnosis. *For the Learning of Mathematics*, 2(2), 31. <https://www.jstor.org/stable/40247736>

- The National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Author. <https://eric.ed.gov/?id=ED226006>
- The NCTM Commission on Standards for School Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. National Council of Teachers of Mathematics.
- Torok, R., & Watson, J. (2000). Development of the concept of statistical variation: An exploratory study. *Mathematics Education Research Journal*, 12(2), 147-169.
<https://doi.org/10.1007/BF03217081>
- Tran, D., & Lee, H. S. (2015). The difference between statistics and mathematics. In *Teaching statistics through data investigations MOOC-Ed*. Friday Institute for Educational Innovation: NC State University.
- Watkins, A. E., Bargagliotti, A., & Franklin, C. (2014). Simulation of the sampling distribution of the mean can mislead. *Journal of Statistics Education*, 22(3).
<https://doi.org/10.1080/10691898.2014.11889716>
- Watson, J. M. (2009). The influence of variation and expectation on the developing awareness of distribution. *Statistics Education Research Journal*, 8(1), 32-61.
<https://doi.org/10.52041/serj.v8i1.456>
- Wild, C. (2006). The concept of distribution. *Statistics Education Research Journal*, 5(2), 10-26.
<https://doi.org/10.52041/serj.v5i2.497>
- Wild, C., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review / Revue Internationale de Statistique*, 67(3), 223-248.
<https://doi.org/10.2307/1403699>

Woods, D. K. (2023). *Transana v5.02* [Computer software]. Spurgeon Woods LLC.

<https://www.transana.com>

Zawojewski, J. S., & Shaughnessy, J. M. (2000). Data and chance. In E. A. Silver & P. A.

Kenney (Eds.), *Results from the seventh mathematics assessment of the National Assessment of Educational Progress* (pp. 235-268). National Council of Teachers of Mathematics.

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.

<https://doi.org/10.52041/serj.v7i2.469>

APPENDIX A

Task A

A large university, North University, reported on their website that 20% of their undergraduate students are intended business majors. Suppose you ask a random sample of 100 undergraduates from North University if they intend to major in business.

1. About how many of the 100 students in your sample do you expect to say they are intended business majors? Why?
2. Suppose your friend also asked a random sample of 100 undergraduates at North University if they intend to major in business. How do you expect the outcome of your sample to compare to your friend's sample?
3. Suppose 24% of the students in your sample said they are business majors.
 - a. Are you surprised by this percentage? Why or why not?
 - b. Does this percentage indicate the information on the website is wrong? Why or why not?
4. What sample outcomes would surprise you? Why?
5. Give a range of percentages of intended business majors that you would expect to get in a random sample of 100 undergraduates from North University. Provide a rationale for the range you give.

APPENDIX B

Task B

We can model this context with a box of plastic beads. In the box, there are several thousand beads, each white or green. Suppose 20% of the beads in the box are white. There is a scooper in the box that allows you to easily pick up 100 beads at one time.

6. Explain how this box of beads models our sampling situation at North University. For example, explain what all of the beads in the box represent, what each of the white beads represent, etc.
7. Predict the outcome of five different random samples of 100 undergraduates from the population of all undergraduates at North University. Record your predictions in the table.

Sample	1	2	3	4	5
Prediction					

- a. Explain how you made your predictions.
 - b. Enter your predictions in the electronic spreadsheet I provide to you.
8. Now, use the scooper to collect and record the outcome of five different random samples of 100.

Sample	1	2	3	4	5
Outcome					

- a. Enter your sample outcomes in the electronic spreadsheet I provide to you.
- b. Why are some or all of these sample outcomes different from each other?
- c. Why are some or all of these sample outcomes different from 20%, the true percentage of white beads in the box?

We will use StatCrunch to help us quickly construct a dot plot of your sample outcomes.

9. How do the actual sample outcomes compare to your predictions?

10. Predict the outcome of 20 different random samples of 100 undergraduates from the population of all undergraduates at North University. Record your predictions in the table.

Sample	1	2	3	4	5	6	7	8	9	10
Prediction										
Sample	11	12	13	14	15	16	17	18	19	20
Prediction										

- Explain how you made your predictions.
- What, if anything, changed in your thinking from your first set of predictions (for the five random samples of 100) to this second set of predictions (for these 20 random samples of 100)?

11. Now, use the scooper to collect and record the outcome of 20 different random samples of 100.

Sample	1	2	3	4	5	6	7	8	9	10
Outcome										
Sample	11	12	13	14	15	16	17	18	19	20
Outcome										

- Why are some or all of these sample outcomes different from each other?
- Why are some or all of these sample outcomes different from 20%, the true percentage of white beads in the box?
- As you were collecting your samples, I entered your predictions and sample outcomes in the electronic spreadsheet. Please verify I entered your predictions and sample outcomes

correctly. We will use StatCrunch to help us quickly construct a dot plot of your sample outcomes.

12. How do these actual sample outcomes compare to your predicted sample outcomes?
13. Would you be surprised if 24 out of a group of 100 randomly selected undergraduates from North University were intended business majors? Explain.
14. Would you be surprised if 38 out of a group of 100 randomly selected undergraduates from North University were intended business majors? Explain.
15. What, if any, outcomes would surprise you? Explain.

APPENDIX C

Task C

To consider taking more samples quickly, we will use an online applet.

16. To get a feel for how this applet works, I will ask you to use the applet to recreate the repeated sampling process you used earlier with the beads. But first, let's explore how the applet works by answering the following questions.

- a. In the applet, the first value to be entered is the probability of success. What is the probability of success in the context of North University undergraduates and how does it relate to the simulation you performed with the beads?
- b. In the applet, the second value to be entered is the sample size. What is the sample size in the context of North University undergraduates and how does it relate to the simulation you performed with the beads?
- c. In the applet, the third value to be entered is the number of samples. How does the number of samples relate to the simulation you performed with the beads?
- d. In the applet, you can choose to display sample outcomes as the number of successes or the proportion of successes. What is the difference between these two?

Earlier, you drew five different random samples of 100 beads from the box of beads that are 20% white, then you reported the number of white beads in each sample.

17. Use the applet to recreate this sampling process. In other words, use the applet to simulate taking five different random samples of 100 undergraduates from North University's undergraduate population which consists of 20% intended business majors. You may choose to draw one sample at a time or all five samples together.

- a. Note the information provided about the most recent results. Explain what this means.

- b. Pick a dot on the dot plot. Explain what this dot means in the context of this problem.
18. Use the applet to recreate the sampling process of drawing 20 different random samples of 100 beads from the box. In other words, use the applet to simulate taking 20 different random samples of 100 undergraduates from North University's undergraduate population which consists of 20% intended business majors. You may choose to draw one sample at a time or all 20 samples together.
- a. What do you notice about the distribution?
19. Use the online applet to simulate taking 500 different samples of 100 undergraduates from North University's undergraduate population which consists of 20% intended business majors.
- a. What do you notice about the distribution?
20. Give a range of proportions you would expect to get in any one single random sample of 100 undergraduates from North University. Provide a rationale for your range.
21. Earlier, you gave a range of _____ and justified your range by saying _____. What changed in your thought process from the first range you gave to this range?
22. Suppose you asked 100 randomly selected undergraduates from North University if they intended to major in business. What outcomes would surprise you? Explain.
23. Take a few minutes to draw a picture that represents the repeated sampling process that the applet is modeling for the North University context. This does not have to be a mathematical or statistical drawing. You may use words, pictures, stick figures, arrows, etc.

APPENDIX D

Task D

Recall from the first interview, North University reported on their website that 20% of the undergraduate students are intended business majors. A nearby large university, South University, does *not* report on their website what percentage of their undergraduate students are intended business majors. You know it is impossible to ask every single undergraduate student at South University what their intended major is, so you decide to ask a random sample of undergraduates instead.

We can model this context using this box of plastic beads. This box is different from the one we used in the first interview in that each bead is yellow or blue. Moreover, we do not know the proportion of the beads that are yellow or blue. There is a scooper in the box that allows you to easily pick up 100 beads at one time. Suppose a yellow bead represents an intended business major at South University.

1. Use the scooper to collect and record one sample outcome. Explain what this outcome represents in the context of the problem.
2. How does this random sample relate to the population of all undergraduate students at South University?
3. Based on this particular sample outcome, do you think the proportion of all undergraduate students at South University that are intended business majors is **the same** as North University (0.20, or 20%)? Explain.
4. Based on your sample, make a prediction for the proportion of all undergraduate students at South University that intend to major in business.

- a. Prediction: _____
 - b. How confident are you in your prediction? Explain.
5. How can you determine if the sample outcome indicates that the proportion of all undergraduate students at South University who intend to major in business is your prediction from question 4a?
6. Perform the simulation you explained in question 5. Please explain your thinking throughout the simulation process.
7. Based on your results of the simulation, how do you feel about your original prediction?
8. What other values do you think would be a good estimate for the proportion of all undergraduates at South University?
9. Explore each of these new values using simulation. Please explain your thinking throughout the simulation process.
10. Give a range of plausible values for the true percentage of intended business majors at South University. In other words, provide a lower bound for the lowest value you expect to be the true percentage and an upper bound for the highest value you expect to be the true percentage.

APPENDIX E

Description of Clips in Interview 1

Clip Number	Prompt Number	Prompt
1	1	About how many of the 100 students in your sample do you expect to say they are intended business majors? Why?
2	2	Suppose your friend also asked a random sample of 100 undergraduates at North University if they intend to major in business. How do you expect the outcome of your sample to compare to your friend's sample?
3	3a	Suppose 24% of the students in your sample said they are business majors. Are you surprised by this percentage? Why or why not?
4	3b	Does this percentage indicate the information on the website is wrong? Why or why not?
5	4, 5	What sample outcomes would surprise you? Why? Give a range of percentages of intended business majors that you would expect to get in a random sample of 100 undergraduates from North University. Provide a rationale for the range you give.
6	7a	Predict the outcome of five different random samples of 100 undergraduates from the population of all undergraduates at North University. Record your predictions in the table. Explain how you made your predictions.
7	8b	Why are some or all of these sample outcomes different from each other?
8	8c	Why are some or all of these sample outcomes different from 20%, the true percentage of white beads in the box?
9	10a	Predict the outcome of 20 different random samples of 100 undergraduates from the population of all undergraduates at North University. Record your predictions in the table. Explain how you made your predictions.
10	11a	Why are some or all of these sample outcomes different from each other?
11	11b	Why are some or all of these sample outcomes different from 20%, the true percentage of white beads in the box?
12	13, 14	Would you be surprised if 24 out of a group of 100 randomly selected undergraduates from North University were intended business majors? Explain. Would you be surprised if 38 out of a group of 100 randomly selected undergraduates from North University were intended business majors? Explain.
13	15	What, if any, outcomes would surprise you? Explain.
14	18	Use the applet to recreate the sampling process of drawing 20 different random samples of 100 beads from the box. In other words, use the applet to simulate taking 20 different random samples of 100 undergraduates from North University's undergraduate population which consists of 20% intended business majors. You may choose to draw one sample at a time or all 20 samples together. What do you notice about the distribution?
15	19	Use the online applet to simulate taking 500 different samples of 100 undergraduates from North University's undergraduate population which consists of 20% intended business majors. What do you notice about the distribution?
16	20, 21	Give a range of proportions you would expect to get in any one single random sample of 100 undergraduates from North University. Provide a rationale for your range. What changed in your thought process from the first range you gave to this range?
17	22	Suppose you asked 100 randomly selected undergraduates from North University if they intended to major in business. What outcomes would surprise you? Explain.
18	23, 24	Take a few minutes to draw a picture that represents the repeated sampling process that the applet is modeling for the North University context. This does not have to be a mathematical or statistical drawing. You may use words, pictures, stick figures, arrows, etc. Please explain your picture.

APPENDIX F

Description of Lorraine's Clips in Interview 2

Clip Number	Prompt Number	Prompt
19	1, 2	Use the scooper to collect and record one sample outcome. Explain what this outcome represents in the context of the problem. How does this random sample relate to the population of all undergraduate students at South University?
20	3	Based on this particular sample outcome, do you think the proportion of all undergraduate students at South University that are intended business majors is the same as North University (0.20, or 20%)? Explain.
21	4	Based on your sample, make a prediction for the proportion of all undergraduate students at South University that intend to major in business. How confident are you in your prediction? Explain.
22	5	How can you determine if the sample outcome indicates that the proportion of all undergraduate students at South University who intend to major in business is your prediction from question 4a?
23	6, 7	Perform the simulation you explained in question 5. Please explain your thinking throughout the simulation process. Based on your results of the simulation, how do you feel about your original prediction?
24	8, 9	What other values do you think would be a good estimate for the proportion of all undergraduates at South University? Explore each of these new values using simulation. Please explain your thinking throughout the simulation process.
25		Based on your sample outcome, what do you think is the lowest plausible value for the true proportion of students at South University who are intended business majors?
26		Based on your sample outcome, what do you think is the highest plausible value for the true proportion of students at South University who are intended business majors?
27		Do you think the true proportion of South University students that are business majors could be 30%? Explain.
28		Use the applet to test your high bound of 60%. Explain your thinking.
29		Do you think the true proportion could be any higher than 60%? Explain.
30	10	Give a range of plausible values for the true percentage of intended business majors at South University. In other words, provide a lower bound for the lowest value you expect to be the true percentage and an upper bound for the highest value you expect to be the true percentage.
31		Please summarize how you used the applet to determine your range of plausible values for the true percentage of intended business majors at South University. What criteria did you use? What were you looking for in order to tell you that you felt confident about your range?