The Role of Sampling Variability in Developing K-8 Preservice Teachers’ Informal Inferential Reasoning

Omar Abu-Ghalyoun

Follow this and additional works at: https://scholarworks.wmich.edu/dissertations

Recommended Citation
https://scholarworks.wmich.edu/dissertations/3433

This Dissertation-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Dissertations by an authorized administrator of ScholarWorks at WMU. For more information, please contact wmu-scholarworks@wmich.edu.
THE ROLE OF SAMPLING VARIABILITY IN DEVELOPING K-8 PRESERVICE TEACHERS' INFORMAL INFERENTIAL REASONING

by

Omar Abu-Ghalyoun

A dissertation submitted to the Graduate College in partial fulfillment of the requirements for the degree of Doctor of Philosophy Mathematics Western Michigan University April 2019

Dissertation Committee:

Mariana Levin, Ph.D., Chair
Christine Browning, Ph.D.
Steven Ziebarth, Ph.D.
Stephanie Casey, Ph.D.
ACKNOWLEDGEMENTS

I first would like to thank my dissertation committee. I am thankful to Dr. Stephanie Casey, my outside reader, for her time and insightful feedback and for inviting me to participate in formative professional opportunities related to my research such as Enhancing Statistics Teacher Education with E-modules (ESTEEM) workshop. I thank Dr. Steven Ziebarth, the third reader, for his untiring support, careful attention to detail, and guidance throughout my journey. I owe an enormous amount of thanks to Dr. Christine Browning, who helped me a lot with collecting my data from her classes and who also acted as the second reader. Her insight and keen advice were invaluable at every step of the research. Most importantly, I thank my advisor, Dr. Mariana Levin, for her guidance, advise, patience, and friendship over the past two years. It wasn’t an easy process, but she advised me, focused me, and challenged me in a successful manner. I feel like a part of the mathematics education research community for her effort. Thanks must also go to all pre-service teachers who participated in this research and whose cooperation and input made this research possible. I wish to express my thanks to my wife, Amani. I could never have accomplished this dissertation without her support and understanding. I also wish to thank my sons Ahmad, Tameem, and Rayan for doing their best to understand a father who had to be confined to his study for such a long time. I am extremely grateful to my parents whose love and guidance are with me in whatever I pursue.

Omar M. Abu-Ghalyoun
THE ROLE OF SAMPLING VARIABILITY IN DEVELOPING K-8 PRESERVICE TEACHERS' INFORMAL INFERENTIAL REASONING

Omar Abu-Ghalyoun, Ph.D.
Western Michigan University, 2019

Recent influential policy reports, such as the Common Core State Standards (CCSS-M, 2010) and Guidelines for Assessment and Instruction in Statistics Education Report, (GAISE, 2007), have called for dramatic changes in the statistics content included in the K-8 curriculum. In particular, students in these grades are now expected to develop Informal Inferential Reasoning (IIR) as a way of preparing them for formal concepts of inferential statistics such as confidence intervals and testing hypotheses. Ben-Zvi, Gil, & Apel, (2007) describe IIR as the cognitive activities involved in informally making statistical inferences. Over this path from informal to formal inference, many important concepts will be integrated into students’ understanding and therefore underpin their IIR ability. One of these fundamental concepts is sampling variability which has been explicitly emphasized in both of the above policy documents. Given this emphasis on sampling variability in the K-8 curriculum, future teachers need support to acquire sufficient content knowledge of this concept. While previous research had outlined general frameworks for what constitutes an understanding of sampling variability (Pfannkuch, 2008; De Vetten, Schoonenboom, Keijzer, & van Oers, 2018), pilot data indicated that a fine-grained analysis of pre-service teachers’ (PSTs) reasoning about sampling variability revealed some facets of reasoning and understanding that were not accounted for in previous frameworks.
This dissertation study investigated a range of non-normative ideas that PSTs employ in reasoning about sampling variability and whether their reasoning processes were sensitive to context. These issues were studied in the context of a content course on statistics and probability for pre-service elementary and middle grades teachers at a midwestern university. Analysis of seven PSTs’ video and screen records of task-based interviews was guided by techniques of Knowledge Analysis (diSessa, Sherin, & Levin, 2016) and identified patterns of non-normative reasoning about four different facets of sampling variability. Identified patterns of reasoning were used to adapt and elaborate Pfannkuch’s (2008) framework for the ways of thinking about sampling variability. More significantly, data analysis also revealed consequential contextualities in how PSTs reasoned about sampling variability in different situations. Implications of this study for teacher education include highlighting the need for using purposefully designed curricula that explicitly emphasize detailed facets of sampling variability across multiple data-contexts.
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS .................................................................................................................. ii

LIST OF TABLES .......................................................................................................................... vii

LIST OF FIGURES ........................................................................................................................ viii

CHAPTER 1: INTRODUCTION AND STATEMENT OF THE PROBLEM ........................................ 1

Introduction .................................................................................................................................... 1

The Importance of Informal Statistical Inference ................................................................. 1

Preparing Future Teachers for Teaching Informal Statistical Inference (ISI) .............. 3

Sampling Variability and Its Role in Supporting Informal Inferential Reasoning (IIR) .... 4

Research Purpose and Questions ............................................................................................ 5

The Significance of the Study .................................................................................................... 6

CHAPTER 2: LITERATURE REVIEW ........................................................................................... 7

The Case of Statistics Education-Recommendations from Policy ....................................... 7

The Place of Informal Statistical Inference in the Curriculum ............................................... 8

Characterizing ISI ...................................................................................................................... 11

Informal Inferential Reasoning ................................................................................................. 13

Sampling Variability .................................................................................................................. 15

Dynamic Statistics Software as a Support for IIR .............................................................. 19

The Role of Context in Developing IIR .................................................................................... 23

Identifying a Focal Studying Phenomenon ............................................................................ 25
Table of Contents—Continued

CHAPTER 3: RESEARCH DESIGN AND METHOD ................................................................. 27

Overview of Research Design .......................................................................................... 27

Identifying a Focal Learning Phenomenon and Research Method ................................. 27

Pilot-Study I: Fall 2017 ..................................................................................................... 28

Pilot-Study II: Spring 2018 ............................................................................................. 31

Dissertation Research Method ......................................................................................... 34

Participants ...................................................................................................................... 34

The Statistics Content Course ........................................................................................ 37

Dissertation Data Sources ............................................................................................... 38

Data Collection Timeline ............................................................................................... 40

Interview and In-Class Tasks .......................................................................................... 42

The Bean Task ................................................................................................................ 45

Voting Task ...................................................................................................................... 46

The Gym Task ................................................................................................................. 46

Chapter Summary .......................................................................................................... 47

CHAPTER 4: THEORETICAL PERSPECTIVE ................................................................ 48

Theoretical Perspective ................................................................................................... 48

Basic Assumptions ........................................................................................................... 49

Illustrative Example of the Use of KiP Framework ......................................................... 53

Chapter Summary .......................................................................................................... 58
CHAPTER 5: EXTENDING A CONCEPTUAL FRAMEWORK FOR SAMPLING VARIABILITY

Introduction ................................................................. 59

Analytical Approach .......................................................... 61

Pfannkuch’s (2008) Framework for Thinking about Sampling Variability ............... 63

Patterns of Reasoning about Sampling Variability Appearing in the Data ............... 65

Summary of the Non-Normative Patterns of Reasoning ..................................... 88

Chapter Summary .............................................................. 90

Looking Ahead to Chapter 6 .................................................. 91

CHAPTER 6: CONTEXTUALITY IN THE CASE OF TANNER ....................... 93

Analytical Framework Used in this Chapter ............................................. 93

Finding and Describing Instances of Contextuality (Analytical Strands Two and Three) ..... 95

Contextuality in Tanner’s Reasoning: Example One ....................................... 95

Contextuality in Tanner’s Reasoning: Example Two ..................................... 105

Summary of Strands Two and Three: Selecting and Describing the Phenomena .......... 112

Re-describing the phenomenon of contextuality using Knowledge in Pieces (Analytical Strand Four) ............................................................. 114

Chapter Summary ........................................................................ 131
### Table of Contents—Continued

**CHAPTER 7: DISCUSSIONS AND IMPLICATIONS** ................................................................. 132

The First Research Question .................................................................................................. 133

The Second Research Question .............................................................................................. 140

Implications for Instruction .................................................................................................... 144

Limitations and Future Work ................................................................................................. 146

Closing Remarks .................................................................................................................. 149

**REFERENCES** .................................................................................................................. 151

**APPENDICES**

A: HSIRB Approval Letter .................................................................................................... 164

B: Participant Consent Form .................................................................................................. 165

C: Beginning of Class Interview Protocol—Fall 2017 .......................................................... 169

D: Beginning of Class Questionnaire—Spring 2018 ............................................................ 172

E: Interview and In-Class Tasks .......................................................................................... 174

F: Interview Protocol Used in Fall 2018 .............................................................................. 179
LIST OF TABLES


3. A table used to organize PSTs’ sampling outcomes during Voting Task.................. 44


5. Four facets of sampling variability along with the non-normative answers............. 67

6. Summary of PSTs answers about different facets of sampling variability................ 88

7. Summary of PSTs non-normative reasoning about different facets of sampling variability.......................................................... 89
LIST OF FIGURES

1. Framework for ways of thinking about sampling variability by Pfannkuch, (2008)…….17
2. Example of hat plots for comparing two data sets using TinkerPlots………………………….21
3. Example of a sampling simulation that generates two sets of data values using TinkerPlots..................................................................................................................................................22
4. TinkerPlots Sampler used in the in-class Voting Task………………………………………………43
5. The population and the scoops used to select the samples in the Bean Task…………………45
6. The sampling distributions provided in the Gym Task…………………………………………46
7. Snapshots of the development of knowledge.......................................................................50
8. A representation of all the knowledge elements cued in Maria’s solution to the Coin Task in Wagner, (2006).......................................................................................................................................................57
9. Recreation of the ten samples selected by Tanner using the small scoop…………………97
10. Recreation of the sampling distribution created during the Bean Task………………106
11. Screenshot of the sampling distribution created during the Voting Task using TinkerPlots.........................................................................................................................................................106
12. Summary of Tanner’s reasoning about two facets of the notion of sampling variability across two learning contexts...............................................................113
13. Tanner’s knowledge system about the relationship between the size of the samples and sampling variability after solving the Bean Task ……………………………116
14. Tanner’s knowledge system about the relationship between the size of the samples and sampling variability after solving the Gym Task.........................................118
15. Tanner’s knowledge system about the relationship between the size of the samples and sampling variability after solving each of the Bean and the Gym Tasks……119
List of Figures–Continued

16. Tanner’s knowledge system after cueing his attention to the equivalence of the Bean and the Gym Tasks ................................................................. 122

17. Tanner’s knowledge system after resolving the disequilibrium about the relationship between the size of the samples and sampling variability ............... 124

18. Tanner’s knowledge system associated with the inverse relationship between sampling variability and the confidence in the drawn inferences in the data-context of the Bean Task ................................................................. 126

19. Tanner’s knowledge system associated with the inverse relationship between sampling variability and the confidence in the drawn inferences after the first response in the gym’s data-context ......................................................... 128

20. Tanner’s knowledge system associated with the inverse relationship between sampling variability and the confidence in the drawn inferences after the second response in the gym’s data context ......................................................... 130

21. Framework for ways of thinking about sampling variability .......................... 139
CHAPTER 1: INTRODUCTION AND STATEMENT OF THE PROBLEM

Introduction

One of the most important goals of today's schools is to prepare students to be proactive thinkers and problem solvers, who are able to manage the complexities of an uncertain and dynamic world. With the substantial growth in technological tools and easy access to information, there has been an increasing interest in incorporating statistics and data analysis into school curricula over the last three decades. In the United States, since the 1980s, school mathematics standards have emphasized the need for developing students’ ability to reason about data and to use them effectively and critically for prediction and decision-making (National Council of Teachers of Mathematics, 1989). Unfortunately, statistics teaching practice in schools has tended to focus more on the computational aspects of statistics (e.g., finding measures of center or making and reading statistical displays) especially in the lower grade levels (Rolka & Bulmer, 2005; Sorto, 2006). Students who are fortunate enough to have some exposure to statistics concepts in high school are typically introduced to formal concepts of inferential statistics such as confidence intervals and testing hypotheses, but without having important foundational knowledge of sampling behavior to make sense of these formal concepts of inferential statistics.

The Importance of Informal Statistical Inference

Statistics, which is also known as the science of learning from data, has two main branches: descriptive statistics and inferential statistics. Descriptive statistics entails the numerical, graphical, or tabular techniques of analyzing and describing a dataset. Inferential statistics, on the other hand, is aimed at making conclusions or estimations that go beyond the
available data based on some observed patterns in the data at hand which might also exist in some broader context.

Traditionally, statistical inference is taught in statistics courses as a set of procedures and formal tests through which the data obtained from the sample is used either to give an estimation for some population parameter (i.e., construct confidence interval), or to test a claim about the value of this parameter (i.e., test a hypothesis). Because of the computational difficulties entailed in formal statistical inferential methods, studying statistical inference has traditionally been postponed until high school or college (Garfield & Ahlgren, 1998).

*Formal statistical inference* involves the use of formal statistical tests based on probability theory. In contrast, *informal statistical inference* (ISI) involves statistical reasoning that has less complexity than formal statistical inference. Pfannkuch (2006) defined the term ISI as “the drawing of conclusions from data that is based mainly on looking at, comparing, and reasoning from distributions of data” (p. 1). Unlike formal statistical inference, ISI does not require the use of formal statistical tests such as testing hypotheses or confidence intervals (Harradine, Batanero, & Rossman, 2011).

With ISI, students might use qualitative descriptions of the behavior of the data instead of basing their assessments on explicit statistical calculations (Makar & Rubin, 2018). Many studies have shown that ISI can be made accessible to elementary and middle school students by harnessing their curiosity about inference and prediction (Meletiou-Mavrotheris & Paparistodemou, 2015; Watson & English, 2016). If students are familiarized with ISI in primary school, this might help them understand the processes involved in formal statistical inference in high school (Bakker & Derry, 2011; Makar, Bakker, & Ben-Zvi, 2011; Browning, Goss, & Smith, 2014). However, this is not the only goal of teaching ISI in elementary school. Many
studies showed that most students and adults do not think statistically about important issues that affect their lives (Garfield & Ben-Zvi, 2007). Teaching ISI in elementary school may help students to deepen their understanding of the way data can support meaning-making about the real-world (Makar, 2016; Makar, Bakker, & Ben-Zvi, 2011; McPhee & Makar, 2014).

Preparation Future Teachers for Teaching Informal Statistical Inference (ISI)

Since the mid-1980’s, there has been a wealth of research on the knowledge that teachers need in order to teach any subject successfully. Shulman (1986) described two broad types of knowledge that teachers need in order to teach successfully: pedagogical content knowledge (PCK) and subject content knowledge (CK). The first type covers (i) the knowledge of content and teaching (KCT) which means the knowledge of how to present, explain and illustrate new material, and (ii) knowledge of content and students (KCS) which includes the knowledge about the students’ conceptions and misconceptions. For CK, teachers need (i) a common content knowledge (CCK) that entails understanding of the subject that students will study, and (ii) a specialized content knowledge (SCK) which includes “how to accurately represent mathematical ideas, provide mathematical explanations for common rules and procedures, and examine and understand unusual solution methods to problems” (Hill, Ball and Schilling, 2008, pp. 377-378).

Given the importance of ISI, some studies investigated K-8 preservice teachers’ (PSTs) content knowledge of ISI (e. g., De Vetten, Schoonenboom, Keijzer, & Van Oers, 2016; Leavy 2010). These studies aimed at evaluating PSTs’ ability to engage successfully with ISI as operationalized by a framework of Makar & Rubin (2007). Makar & Rubin (2007) identify three characteristics of ISI namely: (1) generalizations that are made must extend beyond the data at hand, (2) using prior statistical knowledge to the extent that the knowledge is available, evidence-based justifications for generalizations must be provided; and (3) generalizations must
be described using probabilistic language and explicit reference must be made to the levels of certainty about conclusions drawn.

Each of the components in Makar & Rubin’s framework has been examined in prior research. With respect to the first component—making generalizations that extend beyond available data—De Vetten, Schoonenboom, Keijzer, & van Oers, (2016) reported that most of the participant pre-service teachers in their study tend to describe the data they were reasoning about without understanding that a representative sample can be used to make an inference about a population. With respect to the second component—basing generalizations on evidence-based justifications—De Vetten, Schoonenboom, Keijzer, & van Oers, (2018) found that most PSTs in their study used data as evidence when they compared two samples to generalize to the population. However, the third component— the acknowledgment of uncertainty— has not been explored to the same extent. The studies that exist showed that PSTs have a weak or a superficial knowledge of what causes the uncertainty in the inferences such as sampling variability or sampling bias (e. g., Mooney, Duni, VanMeenen, & Langrall, 2014). Most of the above studies are evaluative in nature in a sense that they evaluate PSTs’ success in meeting the above three components of the ISI as opposed to providing insight into the underlying reasoning processes that supports ISI. This reasoning process is commonly referred to as Informal Inferential Reasoning (IIR).

**Sampling Variability and Its Role in Supporting Informal Inferential Reasoning (IIR)**

Recently, research on IIR has involved a strong emphasis on the concepts of sampling and sampling variability (Ben-Zvi, Makar & Garfield 2017). Pfannkuch, Arnold, & Wild, (2015) argued that learning about the concept of statistical inference should start with some understanding of the nature and behavior of sampling variability. Saldanha & Thompson, (2007)
define sampling variability as “the expectation that samples selected from a population vary among each other and do not exactly match the population, but some aspects might be stable across many samples and therefore these aspects can indicate something about the entire population.” Sampling distributions are recognized as fundamental to statistical inference, and statistics educators suggest that school students should study the informal ideas of sampling distributions (e.g., Garfield & Ben-Zvi, 2007). Pfannkuch et al., (2015) added that, early on, learners should not study sampling distributions as a separate notion, but rather, connect it with the notion of sampling variability as it is associated with the statistical displays. Based on this, if we expect elementary students to be familiar with these notions, sampling variability should be part of future teachers’ specialized content knowledge. However, sampling variability is a complex and multi-faceted concept where true understanding requires learners to be aware of and competently reason with many related statistical ideas (Pfannkuch, 2008). Recent studies indicated that PSTs show a limited understanding of sampling representativeness and sampling variability (Meletiou-Mavrotheris et al., 2014; De Vetten et al., 2016; Mooney, Duni, VanMeenen, & Langrall, 2014).

**Research Purpose and Questions**

Researchers have come to define ISI and operationalize it via frameworks (e.g., Makar & Rubin, 2009; Pfannkuch, 2006). They also have investigated the informal inferential reasoning processes that can lead to ISI and developed theoretical frameworks that capture sampling concepts that underpin this thinking (e.g., Pfannkuch, et al., 2015; Pfannkuch, 2008). However, few studies have investigated PSTs’ inferential reasoning processes, especially as they pertain to the concept of sampling variability that supports this reasoning (e.g., De Vetten, et al., 2018).
Further, the intertwined role of PSTs’ understanding of sampling across multiple contexts and how it could help PSTs in developing IIR has yet to be investigated in fine-grained detail.

Given the lack of research on PSTs’ informal inferential reasoning processes, and in particular, how the understanding of sampling variability is embedded in these processes, the purpose of this study is to (i) characterize PSTs’ reasoning about sampling variability and (ii) to examine the potential existence of contextuality in PSTs’ reasoning processes about sampling variability. In particular, this study addresses the following research questions:

- What non-normative ideas do PSTs invoke in reasoning about facets of sampling variability?
- How are PSTs’ reasoning processes about sampling variability sensitive to data-context?

**The Significance of the Study**

This study addresses sampling variability as a concept that underpins informal inferential reasoning (IIR) which is an important and growing area of research in the field of statistics education. At present, there is limited research about the nature of PSTs' knowledge and reasoning processes about sampling variability that would guide teacher education programs in supporting deep conceptual knowledge of this aspect of the statistics content they will teach. Findings from this study may help the teacher education community to understand in more detail PSTs’ knowledge and reasoning processes related to sampling variability and also recognize the role of context in order to inform instructional approaches used with elementary/middle school PSTs.
CHAPTER 2: LITERATURE REVIEW

In this review of research literature, I describe what is currently known about (i) PSTs’ knowledge of informal inferential reasoning related to sampling variability and how it might be mapped or measured, (ii) computer-based and physical sampling simulations and their role in developing an understanding of the concept of sampling variability, and (iii) the role of context in learning statistics. I set the stage for the discussion of these areas by elaborating upon the recommendations of policy documents related to teaching sampling variability in elementary and middle grades. In doing so, I establish the need for research that examines the ways in which PSTs reason about sampling variability as a building block in their informal inferential reasoning. I also establish the meaning of the key terms used in this study. Because very few studies have investigated PSTs' reasoning about sampling variability, the literature review presented in this chapter draws on both studies with school students and what studies exist to date in the PSTs literature.

The Case of Statistics Education—Recommendations from Policy

As a result of the rapidly increased availability of data, some statistical knowledge has become important for any informed citizen or professional worker in the modern workplace needing to make important choices on the basis of data every day (Franklin et al., 2005). The increased importance of statistics in daily life has resulted in an increased emphasis on school statistics in the US (National Council of Teachers of Mathematics [NCTM], 1989, 2000; Council of Chief State School Officers [CCSSO], 2010; Guidelines for Assessment and Instruction in Statistics Education Report [GAISE], 2007). For example, the National Council of Teachers of Mathematics (NCTM), in its Principles and Standards for School Mathematics (PSSM) (2000), recommended that students start learning statistics as early as kindergarten. This
recommendation to begin to examine statistical ideas as early as kindergarten was also emphasized in the Guidelines for Assessment and Instruction in Statistics Education (GAISE) report which was released by the American Statistical Association’s (ASA) in 2007. However, the GAISE report added more detail to the NCTM Data Analysis and Probability Content Standard and proposed a framework laying out the statistical knowledge that a student should acquire by the end of high school. One of the major recommendations of the GAISE report is viewing school statistics as an investigative process (problem-solving process) in which students try to understand the complexities of real-world situations. The *Common Core State Standards for Mathematics* (CCSSM) (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) also places heavy emphasis on statistics and probability, particularly in grades 6–12, with the GAISE framework having served as a foundation for the development of those statistics standards (Franklin, et al., 2015).

**The Place of Informal Statistical Inference in the Curriculum**

Since the publication of the Curriculum and Evaluation Standards for School Mathematica in 1989 by the National Council of Teachers of Mathematics (NCTM), statistics has found its way into the lower elementary grades. At these grades, students’ exposure to statistical concepts has been primarily limited to descriptive statistics. With respect to statistical inference, the standards included the general statement “make inferences and convincing arguments that are based on data analysis” (p. 105) in grades 5-8.

In 2000, NCTM recommended in the PSSM that the foundations for statistical literacy, reasoning, and thinking, including fundamental ideas of sampling and inferencing, should be taught in the upper elementary and early middle grades rather than being reserved for high school or college courses (National Council of Teachers of Mathematics, 2000).
The GAISE report (Franklin, Kader, Mewborn, Moreno, Peck, Perry, & Scheaffer, 2007) is considered to be the most influential document in the field of statistics education in the 21st century. This report emphasized and elaborated the recommendations of the PSSM regarding statistics education by proposing a two-dimensional framework of processes and levels for the conceptual understanding of statistics in Pre-K-12. Peck, Kader, and Franklin (2008) summarized the key aspects of this two-dimensional framework as shown in Table 1.

Table 1
A framework for conceptual understanding of statistics as appears in Peck et al., (2008)

<table>
<thead>
<tr>
<th>Process Component</th>
<th>Level A</th>
<th>Level B</th>
<th>Level C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formulate question</td>
<td>Questions restricted to classroom</td>
<td>Questions not restricted to classroom</td>
<td>Questions seek generalization</td>
</tr>
<tr>
<td>Collect data</td>
<td>Census of classroom</td>
<td>Non-random sample surveys</td>
<td>Samples designs using random selection</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Begin to discuss the random selection</td>
<td></td>
</tr>
<tr>
<td>Analyze data</td>
<td>Display variability within a group</td>
<td>Quantify variability within a group</td>
<td>Measure variability within a group</td>
</tr>
<tr>
<td></td>
<td>Compare individual to individual</td>
<td>Compare group to group (between) variability in displays</td>
<td>Measure variability between groups</td>
</tr>
<tr>
<td></td>
<td>Compare individual to group</td>
<td>Acknowledge sampling error</td>
<td>Compare group to group using displays and measures of variability</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Some quantification of association</td>
<td>Describe and quantify sampling error based on a simulated sampling distribution</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Understand sampling variability</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Quantification of association</td>
</tr>
</tbody>
</table>
Table 1—Continued

<table>
<thead>
<tr>
<th>Interpret results</th>
<th>Do not look beyond the data</th>
<th>Acknowledge that looking beyond the data is feasible</th>
<th>Look beyond the data in some contexts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No generalization beyond the classroom.</td>
<td>Acknowledge that a sample may or may not be representative for a large population</td>
<td>Generalize from sample to population.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Basic interpretation of models for association</td>
<td>Interpret models for association</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nature of variability</th>
<th>Measurement variability</th>
<th>Sampling variability</th>
<th>Chance variability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Natural variability</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Induced variability</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Focus on variability</th>
<th>Variability within a group</th>
<th>Variability within a group and variability between groups</th>
<th>Variability in model fitting</th>
</tr>
</thead>
</table>

The first dimension consists of the four components of the statistics investigation process, (i) formulate statistical questions; (ii) collect data; (iii) analyze data; and (iv) interpret results, along with two components corresponding to an understanding of the nature and role of variability, (i) nature of variability and (ii) focus on variability. The second dimension provides three hierarchical process levels, A-through-C, on a continuum from novice to proficient for each of the above process components. The goal of the GAISE report is to provide recommendations for all school grade levels where level A tasks are appropriate for elementary students, level B for middle school students, and level C for high school students.
The GAISE report recommended that informal statistical inference (ISI) begins in the elementary grade levels. This recommendation supports the idea of ISI as foundational to the learning of statistics. In process level A, for example, it is emphasized that,

Students also should learn how to use basic statistical tools to analyze the data and make informal inferences in answering the posed questions. Finally, students should develop basic ideas of probability in order to support their later use of probability in drawing inferences at levels B and C (p. 23).

In line with the above recommendations, the Common Core State Standard for Mathematics (CCSSM) recommended that grade seven students engage in “informal comparative inferences about two populations [and use] random samples to draw inferences about a population” (p. 46). Developing ISI at the lower grade levels in US schools is also in line with some international studies that recommended teaching ISI as early as the elementary grade levels through authentic tasks (Makar & Rubin, 2009). In order to translate this recommendation into practice, it is necessary to specify in more detail what might be meant by informal statistical inference.

**Characterizing ISI**

Pfannkuch (2006) described informal statistical inference as “the drawing of conclusions from data that is based mainly on looking at, comparing, and reasoning from distributions of data” (p. 1). Rossman (2008) argued that making inferences necessarily requires going beyond the available data and making generalizations about observed results that reach the population. He also added that making a statistical inference requires a probability model. Makar and Rubin (2009) outlined three features that form a successful informal statistical inference:
1. a statement of generalization that goes beyond the data at hand,
2. the use of the data in hand to support this generalization, and
3. the use of a probabilistic language that indicates some uncertainty about the generalization.

In discussing the first feature, the authors stated that any successful informal inference must go beyond descriptive statistics by making a conclusion about a wider universe (the population) that is beyond the data at hand (a suitable sample). To put the learning of ISI in a curricular context, current U.S. elementary and middle school curricula typically focus only on descriptive statistics topics such as finding the measures of center and variability (both of which pertain to available data only). Making claims about uncertain phenomena based on available data is the most powerful tool in statistics and many statistics educators argue that it should be accessible for students earlier than the formal statistics that students are exposed to in high school courses (Ben-Zvi, Gil, & Apel, 2007; Curcio, 1987; Rossman, 2008).

With respect to the second feature, Makar and Rubin argued that informal statistical inferences need to be grounded in evidence from patterns seen in the data at hand instead of using anecdotes or beliefs about the world. Supporting claims with evidence is essential to developing students’ understandings of the fundamental goal of using statistics. This is unfortunately missed in many schools’ statistics curricula. Instead, the focus has largely remained on creating statistical displays or performing calculations in isolation of the contexts or the purposes from which data came (Friel, Curcio, & Bright, 2001; Pfannkuch, Budgett, Parsonage, & Horring, 2004; Sorto, 2006). When considering the nature of the evidence that is appropriate for informal statistical inference, it is important to keep in mind that the nature of evidence depends on the community in which the evidence is being presented. For example, in
the early grades, observation might be accepted as data in order to promote the initial
development of inferential thinking and reasoning with data. In contrast, in upper grades,
students may be encouraged to critique this type of evidence in preference for more robust and
reliable sources such as testing hypotheses and confidence intervals (Ben-Zvi & Sharett-Amir,
2005).

With respect to the third characteristic of informal statistical inference, the authors stated
that an informal statistical inference should *explicitly express uncertainty in the generalization
statement* because it goes beyond the given data so the inference cannot be stated in deterministic
terms. Elementary and middle grades students, however, will not necessarily quantify this
uncertainty. For example, children might indicate that their claim is just an estimation that
doesn’t apply to all cases.

These three features of informal statistical inference articulated in Makar and Rubin’s
(2009) framework helped the field of statistics education consider how to describe the nature of
informal statistical inference. This has implications for supporting the engagement of elementary
and middle grades students in statistical reasoning (Ben-Zvi & Sharett, 2005; Makar & McPhee,
2009).

**Informal Inferential Reasoning**

Although the above three features characterize ISI, they do not tell us about the reasoning
processes that can lead to ISI, which are commonly referred to as *Informal Inferential Reasoning*
(IIR). Ben-Zvi et al., (2007) define IIR as:

cognitive activities involved in informally drawing conclusions or making
predictions about “some wider universe” from data patterns, data representations,
statistical measures, and models, while attending to the strength and limitations of
the drawn conclusions. Therefore, ISI is just an end product of IIR that doesn’t show the complex reasoning processes (p. 2).

According to Watson (2007), IIR constitutes a continuous experience that starts when learners begin posing statistical questions about data up until they reach the starting point of formal inferential statistics. Over this path from informal to formal inference, many important concepts and ideas will be integrated into a learner’s understanding. Similarly, Rubin, Hammerman, and Konold stated that IIR involves consideration of the following related ideas: (i) properties of aggregates instead of individual cases, (ii) sample size and its influence on the accuracy of estimating the parameters of the population, (iii) controlling bias, and (iv) distinguishing between claims that are always true and claims that are often or sometimes true (Rubin, Hammerman & Konold, 2006). In other words, developing IIR doesn’t necessarily mean that students should learn a specific statistical concept or wait until they are at the appropriate age.

In her synthesis of research on informal inferential reasoning, Reading (2009), named two categories of foundational concepts that need to be mastered in order for students to acquire IIR skills. The first category consists of the following five statistical concepts: (i) variation, (ii) distribution, (iii) the center of data, (iv) the spread of data, and (v) statistical displays. The second category consists of the following four statistical actions: (i) viewing data as aggregates instead of individual cases (cf. Rubin, Hammerman, & Konold, 2006), (ii) accepting proportions rather than absolutes (cf. Ben-Zvi, 2006), (iii) value variability in samples, and (iv) acknowledge randomness as a process. In the following section, I discuss one of the key concepts that underpin the informal inferential reasoning, sampling variability.
**Sampling Variability**

When we make predictions, we usually allow some uncertainty because our predictions are typically based on a sample rather than the whole population. We also judge whether sample data represents its parent population by noticing the patterns in sampling variability. On that basis, any conceptual approach to statistical inference must be built on some robust understandings of the basics of the nature and behavior of sampling variability (Pfannkuch, Arnold, & Wild, 2015). The Common Core State Standard for Mathematics (CCSSM) recommended that grade seven students have experience with generating “multiple samples (or simulated samples) of the same size to gauge the variation in estimates or predictions.” (p. 46).

Given the significant role of sampling variability as a building block for developing the informal inferential reasoning skills, Pfannkuch, et al., (2015) adapted Makar and Rubin’s (2009) framework by incorporating some sampling reasoning concepts that underpin each of the three components in this framework as shown in Table 2.

**Table 2**

*A framework for statistical inference for comparison of two samples of quantitative data by Pfannkuch et al., (2015)*

<table>
<thead>
<tr>
<th>Underpinning sampling reasoning concepts</th>
<th>(Probabilistic) Articulating the uncertainty embedded in an inference.</th>
<th>(Generalization) Make a claim about the aggregate that goes beyond the data at hand.</th>
<th>(Evidence from data) Being explicit about the evidence used, possibly connecting data and context.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling variability</td>
<td>Sample size</td>
<td>Sample Population Distribution</td>
<td>Connecting context</td>
</tr>
<tr>
<td>Uncertainty</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other reasoning concepts</td>
<td>Sampling method</td>
<td>Randomness and Bias</td>
<td>Anything interesting, unusual.</td>
</tr>
</tbody>
</table>
In this framework, they proposed some sampling concepts and ideas that underpin the three components of Makar and Rubin’s (2009) framework. Concerning the probabilistic component, Pfannkuch et al., (2015) stated that learners need to (i) draw on visual imagery of what the shape of the sampling distribution might look like when taking sampling variability into consideration and (ii) draw on their experience of sampling variability to know that they are confident about their claim but not fully certain. For the generalization component, learners need to (i) consider sample distributions, (ii) know that they are making inferences about what exists in the two populations based on samples, and (iii) imagine the shape of the population distributions based on their knowledge of the context (Pfannkuch, Regan, Wild & Horton, 2010). The third component - evidence from data - requires learners to be explicit about the evidence they use based on what they can see in the data at hand. When they compare two sampling distributions, learners might see and point to the cluster in one distribution compared to the other, the position of the averages, or the overall visible spread of the data. To check the reasonability of their claims, they might use some contextual knowledge i.e., connecting evidence to the context.

Another line of research has focused more on sampling variability as a concept, per se. One prominent study was Pfannkuch (2008) in which she adapted Liu and Thompson’s (2007) framework articulating what is entailed in a powerful understanding of probability and used it to describe how grade 10 students developed sampling variability ideas for statistical inference during instruction. Pfannkuch used a web-based sampling simulation with a dynamic visualization tool for quantitative and qualitative data to introduce the ideas of sampling variability. This was done by encouraging the students to notice the variation in the history of sample percentages of one category displayed in bar graphs and the variation in the history of the
sample medians displayed in box plots. Pfannkuch’s data analysis suggested five ways of thinking about different aspects of sampling variability by which students develop an understanding of this concept (see Figure 1). Note that the use of the word “image” in this framework refers to something learners can imagine or conjure up in their minds.

![Diagram of Pfannkuch's framework](image.png)

*Figure 1. Framework for ways of thinking about sampling variability by Pfannkuch, (2008)*

The five ways of students’ thinking identified in this study are associated to the following five facets of sampling variability: (1) the effect of increasing the size or the number of the samples on the appearance of the expected value in the sampling distribution, (2) the effect of the sample size on the location of the expected value in the sampling distribution, (3) the shape of the sampling distribution and how it grows to become symmetric as the number of the selected samples increases, (4) the effect of the sampling method on the sampling outcomes, and (5) the overall purpose of selecting samples which is making an inference about some aspect or characteristic of the parent population.

To preview the analysis and work to come in this dissertation, Pfannkuch’s framework has played a significant role in answering the first research question addressed in this study, *"What non-normative ideas do PSTs invoke in reasoning about facets of sampling variability?"* That is, developing the follow-up and reflection questions included in my interview protocol was informed by this framework. The "images" identified in this framework are similar to the
normative ideas\(^1\) highlighted in my study therefore my data analysis intertwined with these images and provided more elaboration of some of them. More details about these images is provided in chapter five.

The emergence of technological tools that give students easy access to designing and running simulations along with multiple visual representations of data has supported this focus on sampling variability. Most of the research on sampling variability, however, is limited to chance (probability) contexts (e.g., Wagner, 2006; Shaughnessy, 2007). In chance contexts, Shaughnessy (2007) notes that for students to acknowledge sampling variability, they need to develop sufficient understanding of the basics of statistical distributions. This understanding entails changing their tendency to focus on a single expected value and develop intuitions “for a reasonable amount of variation around an expected value” (p. 982). Shaughnessy found some evidence that students’ understanding of sampling variability can be improved if they use physical simulations. In a statistics setting, these findings might also apply, in the sense that collecting multiple random samples of the same size and observing the patterns in the samples’ means or percentages might help students develop intuitions about sampling variability. However, given the small number and size of the samples that can be generated using physical simulations, the outcomes might not clearly show any feature in the data and therefore not help students develop an understanding of sampling variability. Purposefully designed computer-based simulations, that can generate a huge number of simulated samples, along with physical simulations might help them move from “naïve conceptions to richer, more powerful understandings of statistical concepts” (Shaughnessy, 2007, p. 995).

\(^1\) I use “normative reasoning” to indicate reasoning that is statistically accurate and/or appropriate for the context and “non-normative reasoning” to indicate reasoning that is either statistically inaccurate or not applicable to the context at hand.
**Dynamic Statistics Software as a Support for IIR**

Recent advances of educational technologies provide new tools and opportunities for the development of statistical reasoning in the early grade levels. Moreover, some of these tools have the potential to give elementary and middle grades students access to advanced statistical topics (including inferential statistics) and the broader process of statistical investigation that were traditionally postponed for high school courses (Makar & Rubin, 2007). In particular, Dynamic Statistics Software tools have removed some of the tedious calculations that present a barrier to inquiry and have made it easier for school statistics instruction to shift from only learning descriptive statistics procedures (e.g., calculating measures of center) towards more inferential statistics. Recently, several Dynamic Statistics Software have been developed to be used in all grade levels. In particular, two prominent desktop applications are *TinkerPlots® Dynamic Data™ Exploration Software* (Konold & Miller, 2014) and *Fathom® Dynamic Statistics Software* (Finzer, 2006). *TinkerPlots* has been developed primarily to be used by elementary and middle grades students whereas *Fathom* is widely used in secondary schools and colleges for introductory statistics. Because *Fathom* is intended for slightly older learners, it includes more features than *TinkerPlots* does. One more recent development in the field of Dynamic Statistics Software is *Common Online Data Analysis Platform (CODAP)* (Finzer & Damelin, 2016). *CODAP* promises to be a more modern and freely available web-based data platform takes on the needs that drove the development of *Fathom* and *TinkerPlots*. While CODAP is still under development, it shows promise. That is, because CODAP is web-based, it will be more accessible than *TinkerPlots* and *Fathom*. CODAP also makes it easier to share and publish results of data analysis. Because the current study made use of TinkerPlots, I discuss details about the features of this Dynamic Statistics Software in the following sections.
**TinkerPlots**

*TinkerPlots* is a dynamic data-visualization software that has been developed primarily to be used by elementary and middle grades students. One of the fundamental aims of using *TinkerPlots* is to perform genuine data analysis starting with students’ own ideas and moving towards invented statistical graphs and notions (Bakker, 2002). Because of its easy-to-learn interface, *TinkerPlots* can encourage elementary and middle grades students to start exploring data without the need for knowing *a priori* standard types of graphs or of different data types. While doing simple actions such as sorting data into categories, elementary and middle grades students can develop in a bottom-up manner an understanding of both standard statistical displays such as bar graphs or scatterplots (Ben-Zvi, 2000) and how to organize data to make estimations (inferences).

Similar to most other statistics software, *TinkerPlots* also offers computational tools such as measures of center and variability both in numerical and graphical modes. What is special about *TinkerPlots* is that it allows for the user to transform any statistical display into almost any other display easily using some basic and simple actions on data such as order, stack, and separate. A question of educational importance is whether these features and tools that *TinkerPlots* provides are helpful both for students and teachers. Evidence that supports the use of *TinkerPlots* includes a study of Fitzallen that examined *TinkerPlots* and found it satisfied the following six criteria: (i) being accessible and easy to use, (ii) assisting recall of knowledge and representation in multiple forms, (iii) facilitating transfer between mathematical and natural language, (iv) providing extended memory when organizing or reorganizing data, (v) allowing multiple entry points for abstraction of concepts, and (vi) providing visual representations for both interpretation and expression (Fitzallen, 2007, p. 24).
Supporting Statistical Reasoning with *TinkerPlots*. Many studies have been conducted to investigate the effect of using *TinkerPlots* on students’ statistical understanding and reasoning (e.g., English & Watson, 2016; Browning, Goss, & Smith, 2014). In a study focused on the notion of variability, Browning et al., 2014 investigated the development of preservice K-8 teachers’ understanding of the notion of variability in a statistics content course that used *TinkerPlots*. The authors noticed that *TinkerPlots* provided preservice teachers with a “conceptual way of appropriately attending to measures of variability in a manner that the knowledge of procedures could not” (p. 1). TinkerPlots provided the preservice teachers with variety means of thinking about spread with the hat plots and the divider tools. Figure 2 shows an example of hat plots for comparing two data sets using TinkerPlots. The authors conjectured that this would translate into the preservice teachers being better able to support their future students’ development of the concept of variability in the classroom, possibly using TinkerPlots to support that development.

*Figure 2.* Example of hat plots for comparing two data sets using TinkerPlots

**Simulations with TinkerPlots.** *TinkerPlots* includes a simulation tool to define a random experiment by using a “sampler.” The sampler is used by students for modeling chance processes
and for generating simulated samples. For example, students can design a sampler that models collecting samples from some population and then use this sampler to collect and organize data in a table and graphical display for subsequent analysis as shown in Figure 3. This process allows students to experience random sampling and analyze its effect on statistical decision-making. This tool provides a structure where elementary and middle grades students can be introduced to sampling variability and behavior.

Figure 3. Example of a sampling simulation that generates two sets of data values using TinkerPlots

Despite the ease of use and the flexibility of the dynamic visualization tools such as TinkerPlots, some studies revealed that simulating sampling using these tools will not necessarily improve students’ understanding of sampling reasoning (Chance, Ben-Zvi, Garfield & Medina, 2007). Some studies (e.g., Bakker & Frederickson, 2005) found that when students watch simulations of data without designing them or reasoning about what the simulation
represents, they might not believe or understand the outcomes. Therefore, it might be helpful to provide students with physical sampling simulations along with the computer-based simulation.

**The Role of Context in Developing IIR**

An important goal of statistics is to get new knowledge from data and use it to understand some real situation (Wild & Pfannkuch, 1999). The ability to argue about data patterns and suggest new conclusions requires both a strong statistical and a contextual knowledge foundation. Thus, statistics requires both data analysis and social argumentation skills. According to Wild & Pfannkuch, (1999), “the raw materials on which statistical thinking works are statistical knowledge, context knowledge, and the information in data” (p. 228). That necessarily means that statistical knowledge, context knowledge, and the information in data cannot be separated in any successful teaching of statistical thinking (Cobb, 2007).

The term context is usually used to refer to a wide variety of situations including educational circumstances such as classroom settings or real-life stories included in a statistical problem. Therefore, it is important to characterize the context that is used in statistics education and distinguish it from the use of the context in general educational settings. Related to the development of IIR, there are two main types of contexts that should be considered. The first is the student’s learning-experience-context and the second is the data-context (Pfannkuch, 2011).

**Learning-Experience Context.** Learning-experience context is the background knowledge that a student brings to a statistical task in addition to the social and physical learning environment in which they operate. For instance, when a student works on a statistics task that aims at promoting a new learning construction, they bring prior statistical knowledge that stimulates new ideas and concepts through interaction with the teacher, physical and computer tools, and other students in the class.
**Data-context.** Data-context, on the other hand, is the real-world context from which the statistical problem emerged. This context, from a statistical learning perspective, is inextricably tied to solving the problem or gaining more knowledge about the real-world situation (Pfannkuch, 2011). For instance, when analyzing given data in some statistical problem, knowledge of how the data were collected including the design of the study and how variables were defined and measured are considered as parts of the data context.

Wild & Pfannkuch (1999) argued that data-context appears in each stage of a statistical investigation process. For example, when students formulate statistical questions, they reflect on their real-life experience and use it to make the question appropriate. Students are also involved in many context-related activities when they interpret the data during the statistics investigation process such as reflecting on their real-life experience and using it to support their inferences and evaluating the beliefs that they hold about the real world.

Cobb (2007) argued that statistics is one of the hardest school subjects to teach since it’s inextricably tied with data-contexts. In mathematics teaching, however, context might make abstract concepts accessible, but this does not necessarily require the students to consider the context in their answers (delMas, 2004). This means that “statistics requires a different kind of thinking because data are not just numbers, they are numbers with a context” (Cobb & Moore, 1997, p. 801).

Wagner (2006) investigated how a college student transferred pieces of statistical knowledge and skills from one data-context to another. Wagner noticed that the student’s ability to transfer statistical knowledge pieces and skills improved only as the student was able to experience the knowledge and skill in multiple data-contexts. The specific details of how Wagner’s subject came to construe different tasks all involving the Law of Large Numbers as
similar challenged structure-mapping accounts of the process of transfer in the psychological literature (Gentner, 1983). This suggests that in teaching statistics, multiple opportunities should be provided for that engage students in statistical thinking across multiple tasks and data-contexts by focusing on the conceptual similarities underlying the problems involving totally different data-contexts.

Given the fundamental role of the data-context and sampling variability in developing IIR, the current study referred to the following definition of IIR:

IIR refers to cognitive activities [in which students use their current statistical knowledge and data-contexts to (Pfannkuch, 2011)] informally draw conclusions or make predictions about some wider universe from data patterns, data representations, statistical measures and models, while attending to the strength and limitations of the drawn conclusions [by taking sampling variability into account (Pfannkuch et al. (2015)] (Ben-Zvi et al. (2007), p. 2)

**Identifying a Focal Studying Phenomenon**

The previous sections reviewed the literature regarding sampling variability. The review has shown the following: (i) there is a lack of research about the nature of K-8 preservice teachers' knowledge and reasoning processes about sampling variability and (ii) although some conceptual frameworks have been developed recently in order to characterize students' ways of thinking about sampling variability, the intertwined role of data-context in PSTs understanding of sampling variability has yet to be investigated in fine-grained detail. This study will attempt to address these two gaps in the literature through (i) characterizing emergent ways of PSTs reasoning about sampling variability and (ii) examining the existence of the contextuality in PSTs reasoning processes about sampling variability and characterizing it when it exists.
However, this study creates snapshots of PSTs' reasoning about sampling variability and the role of the data-context in this reasoning at one point in time rather than tracks learning over time.
CHAPTER 3: RESEARCH DESIGN AND METHOD

In this chapter, I first address the findings of prior pilot studies and how they influenced the design of this dissertation study. Then I address the research method including an overview of the research design, participants, data sources, data collection timeline, and tasks.

Overview of Research Design

I began by recruiting seven PSTs from one section of a probability and statistics course designed for elementary/middle school teachers. More details about this course are presented later in this chapter. I planned a classroom task and developed task-based clinical interview protocols. A case study was used to identify patterns in PSTs’ knowledge and reasoning processes based upon the video data from the clinical interviews. Recall that the research questions for this study are:

- What non-normative ideas do PSTs invoke in reasoning about facets of sampling variability?
- How are PSTs’ reasoning processes about sampling variability sensitive to data-context?

Identifying a Focal Learning Phenomenon and Research Method

In this section, I address a summary of the findings of the pilot studies and how they, along with the literature review, have influenced my choice of the focal study phenomenon and my choice of the design of this dissertation study which is described in the next chapter.

I had the opportunity to collect pilot data across two semesters before the start of dissertation data collection. The first implementation was in fall 2017 and the second was in spring 2018. My pilot work aimed at testing a recruitment plan and the tasks to be used in the dissertation study, and to focus the research questions according to the findings of the data
analysis in this pilot work. Lessons learned from the first implementation were considered in the next implementation and the lessons learned from the second implementation were considered when conducting the dissertation study in fall 2018.

**Pilot-Study I: Fall 2017**

In fall 2017, all 25 PSTs enrolled in the focal section of the statistics content course gave consent for keeping copies of their work. This included homework, quizzes, exams, projects, etc. Also, four PSTs agreed to be interviewed twice during this semester. Each of these four PSTs was interviewed individually outside the classroom at the beginning of the semester to collect their background data, and their current knowledge about the notion of sampling variability and informal inferences (see Appendix C for the interview protocol used in this iteration). In fall 2017, I “guest” taught the Voting Task (see Appendix E) in class, taking about 35 minutes of class time during week six. This task was purposefully designed to allow for a rich discussion about the fundamental ideas of the notion of sampling variability using a predesigned sampling simulator in *TinkerPlots*. As described in detail in the next chapter, this task discusses both the effect of growing the size of the samples on sampling variability and the effect of growing the number of the samples (with fixed sample size) on sampling variability. One week later, I interviewed two of these four PSTs for the second time—the two other PSTs who I had previously interviewed did not come to the second interview. The goal of the second interview was to get clarification on their written work during the in-class tasks.

**Lessons Learned about the Design of Pilot-Study I.** The following are the lessons learned from the fall 2017 implementation regarding the design and the implementation of the study:
1. Scheduling two interviews with the PSTs was difficult to arrange and so I decided to replace the first interview with an online questionnaire to be sent to all of the PSTs enrolled in the content course early in the semester. The first interview was originally intended to collect some background data and current knowledge about the topic of the study. However, I re-considered that I may be able to collect this data through a questionnaire. I also realized that if I used the identical interview questions on the questionnaire, it would be necessary to adapt and simplify some because they were initially designed for an oral explanation.

2. There was a need for more than 35 minutes for the Voting Task. Because it was the first time the PSTs would be using the sampler feature in TinkerPlots, it took me about 10 minutes to show them how to change the size of the samples, change the number of the collected samples, and how to delete the history. My revised time estimate following this implementation was 50 minutes.

3. The third learned lesson was the need for putting the Voting Task into two separate worksheets and presenting one of them (the growing sample size) in class and using the other (the growing number of samples) during the interview. Teaching the two parts of the Voting Task in sequence during one class meeting was confusing to some of the PSTs because each task required different settings on TinkerPlots. Given that it was the first time for the PSTs to use the sampler feature in TinkerPlots, this transition in settings did pose difficulties for some students.

**Data Analysis for Pilot-Study I.** In the fall 2017 implementation of the Pilot Study, I collected data from the following three sources: (a) the first interview, that aimed at providing some background data about the participant PSTs and their current knowledge about sampling
variability and informal inferential reasoning, (b) written work during the Voting Task, and (c) the second interview which aimed at collecting more detailed data from the PSTs about their written work during the Voting Task.

For the four PSTs who were interviewed early in the semester, I analyzed their interview responses and their written responses during the in-class task. I also analyzed the responses of the two PSTs who came to the second interview. Analyzing the data of the Pilot Study, however, wasn’t intended to go deep into analyzing PSTs’ reasoning about sampling variability, but rather to test the piloted instruments, explore generally how PSTs reasoned about sampling variability and determine whether any interesting phenomena emerged from engaging PSTs in task-based interviews around this topic.

The analysis of the data collected during the fall 2017 implementation of the Pilot Study revealed that PSTs have limited understanding of the notion of sampling variability, given that most of them were confident in the informal inferences they made based on small samples taken from large populations. Participating PSTs were aware of the limitation of the small samples when making predictions about large populations, but they didn’t appreciate collecting many small samples (in spite of the clear cluster of the data that they noticed in the sampling distribution). These findings shifted the initial focus of the study from PSTs’ ability to make successful informal statistical inferences that satisfy the characteristics identified in Makar & Rubins’ (2007) framework toward a more explicit focus on students’ reasoning about sampling variability. In the next implementation, spring 2018, I added some discussion questions to the Voting Task to explicitly focus on the concept of sampling variability.
Pilot-Study II: Spring 2018

In spring 2018, all 16 PSTs who were enrolled in one section of the statistics content course gave consent for keeping copies of their work, including homework, quizzes, exams, projects, etc., to be used as part of the database for the study. Also, four PSTs agreed to be interviewed one time during this semester. Although four PSTs gave consent to be interviewed, only two of them came to the interview. It is difficult to determine exactly why there was this attrition of the participants during both pilots, but my suspicion is that during the time of the semester when the interview took place, the PSTs’ schedules became much busier than what they had anticipated in the first week of class. For the final study, it was decided to provide a small incentive (i.e. a gift card) to the interviewees to encourage volunteering PSTs to remain involved in the study and participate in the single interview.

Data Analysis for Pilot-Study II. In the spring 2018 implementation of the Pilot-Study, I collected data from the following three sources: (a) the online questionnaire which aimed at providing some background data about the participant PSTs and their current knowledge about sampling variability (see Appendix D for the new questions), (b) the task-based interview which aimed at collecting detailed data about their reasoning about sampling variability as they worked on the Voting Task (growing number of samples), and (c) written work during the Voting Task (growing sample size) which was presented in class by the course professor. The focus of the data collection and analysis was on the two PSTs who I interviewed this semester. However, because one of these two PSTs provided richer data during the interview, I decided to focus the analysis on her data. The analysis of the data in this implementation of the pilot was intended to test the Voting Task again and to catch any new
potential phenomena of interest in PSTs’ reasoning processes that suggested further investigating.

In Spring 2018, I interviewed the PSTs before they explored part of the Voting Task in class because I wanted them to come to the interview with a fresh perspective about sampling variability. However, this wasn’t optimal because it took some time during the interview to familiarize them with TinkerPlots. In addition, I reflected that it would actually be better for them to have an introduction to sampling variability in the probability and statistics content course in order to push the discussion during the interview to a deeper level of thinking and reasoning. In the next phase of the study—dissertation data collection—this note was taken into consideration and PSTs were interviewed after they had an introduction to sampling variability and also after experiencing part of the Voting Task (growing sample size) in class.

Similar to the findings from Pilot Study I, the brief analysis of the data collected during Pilot Study II provided more evidence that PSTs might have a limited understanding of the notion of sampling variability in the sense that they seemed confident in the activity about the inferences that they made based on one small sample taken from a large population. The data also revealed that some PSTs became aware that small samples don’t represent the population as a result of observing the differences from one sample to another. Moreover, noticing the cluster of the data around some value in the sampling distribution helped some PSTs trust the repetition of collecting small samples. However, these initial conclusions might not hold in different tasks or data-contexts. That is, we likely can’t argue that these PSTs will give similar answers and use similar reasoning to justify their answers in different tasks or data-contexts. This thought raised the need for probing PSTs’ knowledge and reasoning across multiple data-contexts and using different types of sampling simulations.
In addition to the above design modification, research making use of the Knowledge in Pieces cognitive perspective also made it clear to me that one sampling simulation task within one data-context was not be enough to probe their reasoning about the notion of sampling variability. I selected to focus on sampling variability because there was little prior work on the details of children’s or PSTs’ thinking in this area and thus documenting potentially consequential contextualities could be of particular importance. diSessa (2004) studied college student's learning of the “law of large numbers” in a chance/probability setting and found that contextuality was a dramatic problem for this concept: “systematic integrity is hard won in view of the richness of intuitive perspectives that may be adapted locally to a particular sampling context” (p. 13). In other words, he found that there was a conceptual contextuality that prevented a learner from using the same pattern of reasoning about the law of large numbers in different situations. I thought that in a statistics setting these findings might also apply in the sense that developing intuitions about sampling variability might require a variety of data-contexts.

The development and inclusion of two new tasks, the Bean Task and the Gym Task (described later in this chapter), in the interview were a result of this thinking. These new tasks entail different data-contexts and use different types of sampling simulations—physical sampling simulation. PSTs were required to create sampling distributions in the Bean Task instead of getting ready-made displays using TinkerPlots. PSTs were also required to describe the similarities and differences between these tasks and discuss the effect of the size and number of samples. In the next section, I discuss the research design and method used when conducting the actual dissertation study.
Dissertation Research Method

Participants

I recruited seven PSTs who were enrolled in an elementary/middle school statistics content course in fall 2018 at the same large midwestern university. My plan was to recruit eight PSTs by choosing them from the list of the PSTs who gave consent to participate in the study. The choice of the participants was planned to be based on their performance on four constructed response items released from the Levels of Conceptual Understanding in Statistics (LOCUS) assessment (Jacobbe, 2016). These items were assigned as homework during the first week of the semester in the form of an online questionnaire. However, only seven PSTs gave consent to participate in the study, therefore, I choose all of them regardless of their performance on the LOCUS items. In addition to the interviews, those seven PSTs gave consent to copy their course assignments as well as other written work.

Participants’ Background. Recruitment resulted in seven PSTs, Judy, Alisha, Bella, Emma, Liza, Susan, and Tanner. They were majoring in Elementary education, Early Childhood Education, and Special Education. In addition to the LOCUS items, the online questionnaire contained some questions about their background such as their major and their previous study of statistics. Only Alisha, Susan, Bella had studied statistics in high school. Bella had also taken a statistics course at a community college seven years ago. Their response to the LOCUS items gave an initial indicator of their understanding of the foundational concept of sampling variability at the beginning of the semester. These items entail basic ideas of sampling variability in common data-contexts. In particular, I asked the following two questions:

2 All names are pseudonyms
1. Jon rolled a die (six faces) 10 times and got: 4, 2, 4, 6, 2, 1, 5, 2, 6. He claimed that this die is loaded because #3 has never appeared. Do you agree with Jon? Clearly support your response with all reasons that are relevant to your thinking. Do you have a way by which we can test whether a die is loaded or not?

2. On a particular day, 65% of the births in hospital A were female, and 30% of the births in hospital B were female. On the next day, which hospital is more likely to have more female births than male? Clearly support your response with all reasons that are relevant to your thinking.

3. Imagine the test scores for a group of 1000 middle students in a large school district. The test scores for a random sample of 7 students from this district are 92, 84, 80, 77, 95, 87, 90. Now, consider another random sample of 7 students drawn from this school district. What might be their scores? Predict the 7 values and explain your reasoning.

The first question was intended to probe their intuitive thinking about the effect of the number of the selected samples (number of trials) on the sampling outcomes. Statistically, each rolling of the die represents selecting a random sample of size one and the more we roll the dice the more the sampling outcomes get evened out between the six possible outcomes, therefore less sampling variability. Each of the seven PSTs showed an understanding of the idea that the number of tosses (samples) is not enough and argued that doing more tosses might show the missed number.

The second question was intended to probe their intuitive thinking about the basic idea of making inferences about the population based on selected samples. I expected their responses to include discussion of the trustworthiness of the conclusions made based on a small sample—the percentages of the births in one day only. Only Alisha thought that hospital A is more likely to
have more female births than male on the next day “because it already has a higher chance to obtain more female births than male.”

The rest of the PSTs argued with varying degrees of clarity that it would be difficult to predict which hospital is likely to have more female births the next day based on one-day percentages. This indicates that most of the participating PSTs have some intuitive understanding of the effect of the size of the sample on the confidence of the inference. Bella’s response, given below, was the clearest among the other responses.

Bella: There is not enough evidence to determine which hospital will be more likely to have female births in the future. The data presented are collected from a single day. To be more accurate, we need a sample across a longer time frame.

In the last question, I asked them to predict the scores of students based on a given sample of seven students’ scores taken from a population of size 1000. This question was intended to investigate if the participating PSTs had any intuitive understanding of the idea of expected value. Only Bella was aware that the given sample is not enough to predict the new scores because the population is big compared to the given sample. The rest of the PSTs inferred without hesitation that the new scores would be within the same range of the given seven scores. Their responses to this question may indicate that most of the participating PSTs were not aware of long-term patterns in data or the expected value at this early point of the semester.

In summary, their responses to the LOCUS items demonstrated that the participating PSTs had some intuitive understanding of the overall purpose of the selecting samples which is making conclusions about the parent populations. They also seemed to understand that larger samples are more representative of the population. However, it wasn’t surprising that they were not aware of other facets of the notion of sampling variability such as the law of the large
numbers and the expected value because “the notion that sampling variability decreases in proportion to sample size [the law of the large numbers] is apparently not part of man's repertoire of intuitions” (Kahneman & Tversky, 1972, p. 444). Moreover, their responses revealed that the small sample of PSTs was fairly homogeneous in the sense that they had a similar understanding of the notion of sampling variability, except for Bella who had taken a statistics class at a community college seven years ago.

The Statistics Content Course

The elementary/middle school probability and statistics course is one of three content courses offered by the Department of Mathematics that are required of all students in the elementary education program. The 15-week course meets for two 100-min sessions per week and attends to concepts of statistics and probability appropriate for elementary and middle school teachers. Topics include the statistical investigative process involving formulating questions, techniques for organizing, presenting, analyzing, summarizing, and interpreting data; chance; simulation methods; and analytic methods in probability.

In the fall 2018 semester, two sections of this course were offered with all participating PSTs in this study enrolled in one of these two sections. Because authentic data-contexts may scaffold students’ development of statistical inferential reasoning by providing language supports for talking about statistical ideas (Makar & Confrey, 2005), PSTs in these classes participated in many data-centered activities in contexts familiar to them. The PSTs collected, explored and analyzed data using physical and computer-based sampling simulations, then formulated and evaluated data-based inferences.
Dissertation Data Sources

Data collected for this study came from video recordings and written responses during task-based clinical interviews (e.g., Creswell 2007; Goldin, 1997), and from videos of both whole- and small-group discussions during one session of the statistics content course during the fall 2018 semester. Other data were drawn from the screen recording of the laptops of the participating PSTs. The purpose of using these screen recorders was to capture their work on TinkerPlots because the single class video camera couldn’t capture all of the PSTs’ laptop screens simultaneously. Although the interviews were video recorded, PSTs were sitting facing the camera, thus screen recorders were also used so that screen data could also be captured for analysis.

Classroom Data. In week five, I collected data related to a focal task on the effect of the size of the sample on the stability of sampling outcomes. In this task, PSTs used TinkerPlots to simulate collecting samples of growing size. To capture detailed data only from the PSTs who would be interviewed later, I used two standing cameras to video record their discussion with each other and with the instructor. Six of the participating PSTs were sitting in two groups that had the available cameras. The seventh participant was in a third group. I did not have the opportunity to gather video data during class on this PST, but I was able to audio record her conversation with groupmates and record the screen of her laptop. While the course instructor facilitated the class work on this task, I was on hand to clarify anything for the PSTs in my sample without interfering in the teaching process, in particular, how to use the screen recording software. Early in the semester, I sent each of them an email asking them to download a free and open source screen recording software, OBS Studio, on their laptops. I included in this email the link to the downloading website along with some instructions on how to use it. This screen
recorder was turned on while they were working on the in-class task also during the interview. The use of OBS Studio, which is free and open-source, is important because it provided me access to the different samples that they collected during the simulation, as well as access to an audio recording of their group conversation.

**Task-Based Clinical Interview Data.** In week six in the fall 2018 semester, I interviewed the seven PSTs in my sample, each for approximately 60 minutes. The task-based clinical interviews took place in a quiet room in a building on the main campus of the university. As mentioned in the data collection timeline below, the interviews were conducted in weeks six and seven of the semester because I expected that I may be able to observe a wider range of reasoning patterns, including non-normative\(^3\) patterns of reasoning, about sampling variability at this early time of the semester. During the interviews, I asked the participating PSTs to work on three tasks that involved computer-based and physical sampling simulations, namely the Bean Task, Voting Task (growing number of samples), and the Gym Task. Both the Bean Task and the Gym Task were new to the PSTs therefore they allowed for invoking new reasoning. It is important to stress that the interviews were not focused on teaching new content to the PSTs but rather to provide a means for new patterns of reasoning to emerge that could be analyzed. Therefore, this study creates snapshots of PSTs' reasoning about sampling variability and the role of the data-context in this reasoning at one point in time rather than tracking learning over time. In order to have access to details about the nature of the knowledge and reasoning processes PSTs used as they reason about sampling variability across different data-contexts, I asked them to justify their thinking as they worked on the interview tasks. I developed my planned

---

\(^3\) Recall that I use “normative reasoning” to indicate reasoning that is statistically accurate and/or appropriate for the context and “non-normative reasoning” to indicate reasoning that is either statistically inaccurate or not applicable to the context at hand.
interactions and interview questions carefully based on the feedback that I got from the implementation of the Pilot Studies in fall 2017 and spring 2018, and also based on the feedback that I got from my colleagues and professors during two meetings of a mathematics education seminar. For this feedback, see the interview protocol in Appendix F.

Data Collection Timeline

The following is a summary of the timeline for the collected data:

- **Week One:** Class began August 29th and during the first week, the instructor assigned a background survey that included initial thoughts about sampling and IIR as a homework task. This survey contained the four constructed response items released from the LOCUS assessment (Jacobbe, 2016) as discussed above.

- **Weeks One and Two:** During the first two weeks, PSTs were introduced to the basic concepts of probability including the probability continuum. They also were introduced to the notion of randomness as the lack of pattern or predictability in events rather than haphazard. In particular, the class discussed some features of randomness such as more than one possible outcome; unpredictable outcomes in the short run but a regular and predictable pattern in the long run; uncontrolled sequences of outcomes; and many repeatable trials. The goal of introducing these basic concepts of probability in these early weeks was to familiarize the PSTs with the probability language that they would need when they studied sampling methods in the upcoming weeks of the semester.

- **Week Three:** The researcher obtained signed consent forms from the participating PSTs.

- **Week Four:** PSTs continued studying the statistical investigative process: developing a good statistical question; collecting and looking at data; what is variability? For “collecting and looking at data,” PSTs were introduced to the difference between random
sampling (an unbiased way to select a representative sample) and non-random sampling, such as convenience or volunteer sampling. They also studied features and issues of different types of sampling, including random, systematic, convenience and volunteer. In preparation for the in-class Voting Task (see Appendix E), participating PSTs were asked to download the OBS Studio screen recorder.

- **Week Five:** PSTs continued studying the statistical investigative process: displaying data; various data types. As part of “displaying data,” they studied an introduction to sampling distribution through an activity in which they investigated the length of the words in Gettysburg Address by selecting words randomly from this address and developing a sampling distribution. The discussion in class was mainly focused on the meaning of collecting a random sample. As they displayed the means of their samples, PSTs could notice the variability in sample means, so this helped set the stage for sampling variability. Although the class had not yet studied measures of center or variability, ideas about center came up with finding the average length of the words in Gettysburg Address. Rather than pausing to carefully define what was meant by average, PSTs used the mean (typically defined by them as “the average.”) The Voting Task (growing sample size; see Appendix E) was taught in class this week. In this task, PSTs were introduced to the TinkerPlots sampler tool which they used to investigate the effect of the growing size of the sample on sampling variability.

- **Weeks Six and Seven:** Task-based interviews were held in weeks six and seven. A third party transcribed all of the interviews then I reviewed and iteratively improved them. Course content during weeks seven through fourteen is not directly related to the research questions of this study so there was no data collected in those weeks.
Interview and In-Class Tasks

Zieffler, Garfield, delMas, & Reading, (2008) categorized the tasks that have been used in the research studies about IIR into the following three general types:

1. Using sample data to reason about characteristics of a population. (Ben-Zvi, 2006; Pratt, Johnson-Wilder, Ainley, & Mason, 2008; Zieffler et al., 2008);

2. Comparing samples of data to reason about possible differences between the populations from which they were sampled. (Makar & Confrey, 2002; Makar & Rubin, 2007; Pfannkuch, 2006; Watson & Moritz, 1999); and

3. Judge which of two competing claims is true based on a sample of data (Stohl & Tarr, 2002; Tarr, Lee, & Rider, 2006).

Since this study focuses on sampling variability as a construct that supports the development of IIR, the four tasks that I used in this study fall within the first category of tasks described above. That is, in each of the tasks used in this study, PSTs were asked to collect samples and use them to make inferences about some characteristic of the parent population. In the following section, I describe each of the four tasks (See Appendix E for complete copies). These tasks were: (1) Voting Task (growing sample size), (2) Bean Task, (3) Voting Task (growing number of samples), and (4) Gym Task. (Listed in the order of use in the study.)

In-class Voting Task (growing sample size)\(^4\)

The goal of this in-class task was to investigate the effect of the size of the samples on the sampling variability. PSTs were asked to use a pre-designed TinkerPlots’ sampler that simulated collecting samples from California state voters who were asked whether they would vote for or against a proposition (hereafter referred to as Prop 223). PSTs were asked to simulate the

\(^4\) The Voting Task was adapted from Sowder, Sowder, and Nickerson, 2014
selection of samples of different sizes and notice the percentage of people who voted for Prop 223 across all of the samples. They were to determine what they believed to be the likely outcome of the votes on Prop 223 and explain their thinking.

I designed a TinkerPlots sampler so that the percentage of the voters who would vote for Prop 223 was set as 63% (this percent was also assumed in the original source of this task). Design features of the sampler allowed me to hide this information from the PSTs, making the sampler be more “true to life” for collecting sample data from a population; we typically do not know exactly how people will vote on the day of the election. The sampler, as shown in Figure 4, had a table and graphical display for the immediate sample (the upper two boxes), and a tabular record with a graphical display for the history of the collected samples (the lower two boxes).

![Figure 4. TinkerPlots Sampler used in the in-class Voting Task](image)

While they worked in groups in class, each PST was asked to describe (write down) what they noticed in the sampling distribution as the size of the samples grew and to make inferences about the percent of the voters who would vote for this proposition based on the sample data they
collected. Each member of the group chose one sample-size option from Table 3 below (for example, Person 1: 30 samples each of size 100) and then entered their results into the relevant column. With the increase of the size of the samples, PSTs were expected to notice the decreased variability in the percent of the people who voted for Prop. 223 and also notice the cluster of the sampling data around the row of 63%-65%. The data in the table is expected to take the shape of an isosceles triangle with the first column having more variability (the base of the triangle) and the last column having the least, with the values centered around the 63%-65% row and the rest of the data symmetrically distributed on both sides.

Table 3
_A table used to organize PSTs’ sampling outcomes during Voting Task (growing sample size)_

<table>
<thead>
<tr>
<th>Percentages of the people who would vote for Prop. 223 in the samples</th>
<th>Person 1: 30 samples each of size 100</th>
<th>Person 2: 30 samples each of size 500</th>
<th>Person 3: 30 samples each of size 1000</th>
<th>Person 4: 30 samples each of size 2000</th>
<th>Person 5: 30 samples each of size 3000</th>
</tr>
</thead>
<tbody>
<tr>
<td>48%-50%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>51%-53%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>54%-56%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>57%-59%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60%-62%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>63%-65%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>66%-68%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>69%-71%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>72%-74%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75%-77%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The Bean Task (used during the interview)

This task required the participating PSTs to provide reasoning about the fundamental ideas of sampling variability in a new data-context. The goal of this task was to estimate the percentage of red beans in a container that held red and white beans as shown in Figure 5.

Figure 5. The population and the scoops used to select the samples in the Bean Task

The actual percent of the red beans in the container was set at 20%. PSTs were asked to scoop up samples using two scoops of different sizes, develop sampling distributions, and then use their displays to draw inferences about the number of the red beans in the container (the population). They also discussed the effect of using the larger scoop instead of the small one on the shape of the sampling distribution and the confidence of their inferences using each size of the scoop. See the Bean Task and the associated follow-up questions in the Interview Protocol in Appendix F.
**Voting Task** (growing number of samples, used during the interview)

Similar to the Voting Task (growing sample size) used in class, this interview task included the use of the same *TinkerPlots* sampler, however, this time the PSTs collected multiple samples of the *same size* (n=15). The size of the samples doesn’t change in this task but the number of these small samples grows to the extent that a cluster appears in the sampling distribution. Each PST was asked to make many inferences about the percentage of voters who would vote for the proposition based on different numbers of samples of size 15. The data clustered around the value of 63% in the sampling history display (sampling distribution) as they generated more samples. See Voting Task (growing number of samples) and the associated discussion questions used during the interview in Appendix F.

**The Gym Task (used during the interview)**

Similar to the previous tasks, the Gym Task involved a new data-context and required reasoning about the notion of sampling variability. In this new data-context, the question “What is the typical time spent at the gym?” was investigated by selecting many samples from a population of 800 gym members. Two different sampling distributions (dot plots) of sample means calculated from random samples of the population were given as shown in Figure 6 below.

*Figure 6. The sampling distributions provided in the Gym Task*
These two sampling distributions are different in their spread and ranges. The discussion questions focused on comparing the shapes of these two sampling distributions and what might have caused these differences including the size of the selected samples in each case. Also, the reliability of each sampling distribution for making inferences about the parent population was discussed. Through each of these tasks, PSTs were encouraged to make multiple arguments and externalize their thinking about the effect of the sample size and the number of the selected samples on the variability of the sampling outcomes and on the confidence of the drawn inferences. Moreover, they were asked about the similarities and differences between these tasks which entail sampling variability but in different data-contexts.

**Chapter Summary**

In summary, this study was conducted in the context of an elementary and middle grades PSTs course focusing on probability and statistics. Data sources were task based clinical interviews with seven PSTs and a background survey that elicited PSTs’ initial thoughts about sampling and IIR. In this chapter, I presented the findings of prior Pilot Studies and how they influenced the design of this dissertation study. Then I addressed the research method including an overview of the research design, participants, data sources, data collection timeline, and tasks. The theoretical framework and prospective that guided data analysis in the subsequent chapters is the focus of the next chapter.
CHAPTER 4: THEORETICAL PERSPECTIVE

In this chapter, I first address the theoretical perspective employed in this dissertation study. I then illustrate the use of key constructs of the theoretical framework that are relevant to my work by presenting an example of a study that used this framework.

Theoretical Perspective

The analysis of the data in this study was informed by an epistemological perspective referred to as Knowledge in Pieces (KiP) (diSessa, 1988, 1993). KiP has been used by mathematics educators to examine emerging competence in the domains of fractions (Smith, 1995), probability (Wagner, 2006), calculus (Jones, 2013), and algebra (Levin, 2018). KiP has characteristics that made it a useful candidate as a foundational approach for this study. Specifically, its fine-grained quality suits the questions of the present research, allowing a productive examination of processes of knowledge re-organization. In doing so, it enables me to zoom in on the learning process and analyze the cognitive dynamics of the transitions that occur in knowledge reorganization.

KiP has its roots in studies of students’ reasoning about the physical world (diSessa, 1993) with many aspects of the theoretical perspective that are most apparently related to the case of reasoning about physics. That said, the program of work outlined by diSessa has implications for studying knowledge and learning processes more broadly and thus, in using and adapting this framework to study statistical reasoning, I am contributing to the growing body of work that develops the perspective beyond its origins in physics reasoning (see diSessa, Sherin & Levin, 2016 for a review).
Basic Assumptions

One of the central ideas of the Knowledge in Pieces (KiP) perspective is that knowledge can be productively modeled as a system of diverse knowledge elements that are abstracted from experience. In particular, students' intuitive ideas are considered to be potentially productive resources — neither correct nor incorrect in and of themselves — from which more coherent and integrated knowledge systems can be developed. Knowledge systems, therefore, are considered to be made up of numerous elements of knowledge each of which is not necessarily right or wrong in isolation, but as productive or not for a particular context. This view thus offers a different way of conceptualizing “misconceptions” (See Smith, diSessa & Roschelle, 1993; Brown, Danish, Levin & diSessa, 2016).

As individuals reason about new situations, they activate their prior knowledge elements. From the KiP perspective, learning is modeled largely as a process of reorganization in which, through feedback of various kinds, the use of existing knowledge elements is improved over time to better fit the context. This is often referred to as “tuning [the knowledge system] towards expertise” (diSessa, 1993). More broadly, learning a new idea can be considered as a transformation of one knowledge system into another. This process of knowledge transformation contrasts with replacement models in that there may be many common knowledge elements between the knowledge systems of beginners to a domain and experts. Yet for the experts, the same knowledge elements may be activated in more refined (and contextually appropriate) ways.

Figure 7 gives a schematic representation of the above ideas related to the transformation over time from naïve to conceptually competent. It is offered as a heuristic tool in thinking about “understanding” as a phenomenon of a knowledge system (as opposed to a framework to be
applied, per se). The geometric shapes in the figures are intended to communicate knowledge elements of diverse types. In the “naïve” snapshot, the connections between knowledge elements are tenuous, whereas the connections between knowledge elements in the conceptually competent snapshot are stronger. One can see that some of the same elements that existed even in the naïve knowledge system still play a role (albeit different) in the conceptually competent snapshot. An example clarifying the constructs of the KiP framework follows this section.

![Figure 7. Snapshots of the development of knowledge](image)

In line with Piagetian constructivism (e.g., von Glasersfeld, 1991; Confrey, 1986) the KiP perspective assumes that different contexts can be interpreted through different combinations of knowledge elements (i.e., schemas). Therefore, lack of systematicity across different contexts might prevent learners from noticing differences (contradictions), and the sensitivity of the contexts might prevent them from seeing similarities. Thus, within this view, the role of context is significant. That is, some ideas (knowledge elements) be activated and used in some contexts, and not in others. Furthermore, as a learner’s knowledge system becomes more well-organized, the contexts in which these knowledge elements get consistently activated may change.

---

Understanding how knowledge activation and use depends on context is a core phenomenon KiP aims to capture.

**More specific constructs.** According to diSessa (1993), a fundamental mechanism that affects the dynamics of the knowledge systems is the *activation* and *deactivation* of existing knowledge elements. These dynamics of knowledge systems can be described in terms of *cueing priority* which is “the degree to which a particular knowledge element’s transition to an active state is affected by other previously activated elements,” and *reliability priority* which is “potential feedback that can reinforce or undo the initial activation” (p. 313).

Within the KiP epistemological perspective, *coordination classes* (diSessa & Sherin, 1998; diSessa & Wagner, 2005; diSessa, Sherin, & Levin, 2016) are knowledge systems that model cognitive structures useful for describing particular types of concepts in physics and mathematics such as force, expected value, rate of change, etc. The coordination class model mainly addresses the difficulty that individuals face when they learn concepts that they must operate or recognize in several contexts, especially those in which they are likely to need quite different strategies to operate the concepts across contexts.

DiSessa and Wagner (diSessa, 2004; diSessa; Wagner, 2005 & Wagner 2010) further elaborated this model by calling the set of knowledge elements and reasoning strategies that enable the learner to recognize and apply the concept within some context as the *concept projection*. The range of contexts across which the learner’s concept projections are found to be applicable constitutes the *span* of the concept projections. Thus, a learner’s understanding of some concept might be supported by a variety of concept projections, which in turn might consist of many knowledge elements shared by different concept projections. DiSessa and Sherin (1998) called the learner’s ability to recognize and apply the same concept projections across different
contexts as *alignment*. However, if the learner sees contrary indications of some cued knowledge projections associated with the same concept, the degree to which they experience any *disequilibrium* would depend on their relative confidence in each of the contrasting knowledge projections (Izsák and Jacobson, 2017). If they had high confidence in each of the contrasting knowledge projections, their experience of disequilibrium might be strong and difficult to resolve. If they had more confidence in some knowledge projections than others, their experience of disequilibrium might be resolved more easily in favor of the cued knowledge projections that provided more sense to them. Having a fully developed coordination class entails a learner’s ability to (a) integrate all of the relevant information in a particular context and (b) aligning the different concept projections across the range of applicable contexts. Thus, a fully developed coordination class reflects expertise.

Using the notion of coordination classes and particularly the ideas of the concept projections, Wagner, (2006) offered an alternative explanation of *knowledge transfer*. The canonical explanations in the psychological literature for transfer involve mechanisms such as structure mapping, involving a subject abstracting a context-independent knowledge structure from one context and then transferring their understanding to a new context. In contrast, Wagner’s account describes the process of transfer as “incremental growth, systematization, and organization of knowledge elements that only gradually extend the span of situations in which a concept is perceived as applicable” (p. 10). Because of its closeness in terms of both the domain of study (chance and probability reasoning) and constructs Wagner’s used, I provide more details about this study in the next section.
Illustrative Example of the Use of KiP Framework

Below I illustrate the use of key constructs of the KiP framework that are relevant for my work by presenting an example of a study that used this framework. I chose Wagner (2006) as an example because his analysis focuses on a similar kind of learning phenomenon (contextuality of learners’ understanding) with a related topic (Law of Large Numbers).

Maria, the subject in Wagner’s study, was a 21-year-old freshman Mass Communications major enrolled in a required statistics course. Wagner interviewed the subject many times during the semester and used many tasks associated with the law of large numbers. The first task was the “Coin Task.” The main question asked of Maria in this task was, do you choose a small or large number of trials if you would like the percentage of the heads tosses to be around 70%? Based on the law of large numbers, Maria should choose a smaller number of trials to get a percentage of around 70% because a very large number of trials will most likely give her 50% heads. In the earliest stages of Maria’s reasoning about the Coin Task, she had a primitive understanding of the expected value of a sequence of flips of a coin, but she did not yet have the ability to coordinate the expected value with the experimental outcome of a series of flips which is key to understanding the law of large numbers. A month later, Maria was interviewed and asked again about the Coin Task. This time she managed to answer the question and discuss the relevance of the law of large numbers. Wagner, however, argued that Maria had not yet constructed a concept projection for the expected value that would enable her to recognize it as relevant to another task.

After discussing the Coin Task, the interviewer (W) provided Maria (M) with a paper circle spinner made up of 10 equal angular sectors. Seven of the sectors were blue, and three were green. Wagner then asked Maria whether she would want a less or greater number of spins
if the goal was to get the percentage of the blues to be between 40% and 60% of the total number of spins.

M: OK. … Land on blue? … Well, 70% of that circle is blue. Yeah. Seventy percent of it is blue, so, for it to land between 40 and 60 percent on blue, then, I would say there really is no difference. [She means it doesn’t make a difference whether one does few or a lot of spins.]

Jo: Why?

M: Because if 70% of the circle, or, yeah, the spinner is blue, so … it’s most likely going to land in a blue area, regardless of how many times I spin it. It kinda really doesn’t matter. It’s not like the coins (Wagner, 2006, p. 24).

According to Wagner’s analysis, the above answer indicates that Maria does not see the Spinner Task as a situation in which the law of large number works, but she seems to be aware that in any spin, there is 70% chance of getting blue, and 30% chance of getting green. For individual spins, Maria’s reasoning about “chances” is accurate. But with the long-term mean, she did not think that the “expected percentage” of blues existed. She could easily see and argue successfully about the probability of each color in individual spins, but she couldn’t extend this idea to predict a long-term mean, nor did she know if a long-term mean existed in the spinner case at all. We see evidence in her response, “It’s not like the coins,” that the knowledge elements she used to interpret the Coin Task were cued or available for comparison in this new data-context. However, for her, these knowledge elements did not apply in this new data-context. In the language of the KiP framework, Maria did not see the spinner situation as essentially
similar to the Coin Task so couldn’t align the activated concept projections cued in both tasks (data-contexts).

After that, Maria was introduced to a computer-based simulation of the spinner that provided histogram displays for the percents of blue in the sampling outcomes. Before running the simulator, Wagner asked her if she expected the percentages to cluster around any value. Maria was hesitant to give any answer at all. But she reluctantly said that the values might cluster around 70%. When Wagner ran the simulation 20 times, Maria exclaimed, “Wow! It does peak [cluster] at seventy!” (p. 36). She showed clear surprise and pleasure because the simulation confirmed her previous prediction. By virtue of the computer-based simulation, the case ended quickly, which Wagner argued was instrumental in convincing Maria that the expected value also existed in the spinner case. Wagner argued that what was important in understanding Maria’s reasoning was how she perceived the situations, Spinner and Coin, differently. This difference in perception led to apparent contextuality in reasoning across tasks.

The last case of contextuality in Wagner’s study occurred when Maria was solving the Post Office Task as shown in Table 4. When Maria was asked to answer the Post Office Task she struggled at first and couldn’t give any answer. She then hesitantly inferred that large sets of numbers have smaller means. Again, the law of large numbers was not evident to her in this task (data-context). To help her answer the question, Wagner presented a simpler question that involved a new data-context. He asked, would you rather choose a large or small sample of men at a university in order to find the mean height? Quickly and confidently Maria answered a larger sample would be more representative. Wagner argued that the new question cued (helped Maria recognize as being relevant) the knowledge element that larger samples are more representative which had not been mentioned (cued) in any of the previous data-contexts. Once cued, Maria
constructed a concept projection that enabled her to recognize and apply this knowledge element productively in the Post Office data-context (the task Wagner had asked her initially).

Table 4
*The Post Office Task as adapted in Wagner, (2006)*

| Table 4 The Post Office Task as adapted in Wagner, (2006)*
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>When they turn 18, American males must register for the draft at the local post office. In addition to other information, the height of each male is recorded. The national average height of 18-year-old males is 5 feet 9 inches.</strong></td>
</tr>
<tr>
<td><strong>Every day for 1 year, 25 men registered at Post Office A and 100 men registered at Post Office B. At the end of each day, a clerk at each post office computed and recorded the average height of the men who had registered there that day.</strong></td>
</tr>
<tr>
<td><strong>Which would you expect to be true? (circle one)</strong></td>
</tr>
<tr>
<td>1. The number of days on which the average height was 6 feet or more was greater for Post Office A than for Post Office B.</td>
</tr>
<tr>
<td>2. The number of days on which the average height was 6 feet or more was greater for Post Office B than for Post Office A.</td>
</tr>
<tr>
<td>3. There is no reason to think that the number of days on which the average height was 6 feet or more was greater for one post office than for the other.</td>
</tr>
<tr>
<td>4. It is not possible to answer this question.</td>
</tr>
</tbody>
</table>

As the above episode shows, some knowledge elements or concept projections—even if they are not usually cued—can be helpful when brought to the individual’s attention in one particular data-context because they might be combined and coordinated; therefore, they are more likely to be used productively in an increasingly wider span of data-contexts. The combination of *productivity of knowledge* and *contextuality of knowledge use* in Wagner’s analysis illustrates a key feature of analyses informed by KiP.

---

Wagner (2006) inferred that learners often face alignment difficulties as they learn new concepts, for example, the law of large numbers. Some learners might need an exceedingly wide span of different data-contexts for the uses of the law of large numbers before they develop expertise (a well-formed coordination class) for this concept because thinking in different contexts involves constructing concept projections that may need to be learned separately for different data-contexts. Through his data analysis, Wagner developed graphical representations of the knowledge elements cued in each of the used data-contexts. Figure 8, for example, shows the knowledge elements determined by Wagner’s analysis as being cued in the Coin Task. These graphical representations served as inspiration for some of the representations of reasoning of the PSTs in my analysis and thus I present them here to contextualize later aspects of my analysis.

![Phenomena involving probabilities](image)

*Figure 8. A representation of all the knowledge elements cued in Maria’s solution to the Coin Task in Wagner, (2006)*
Chapter Summary

KiP has characteristics that made it a useful candidate as a heuristic epistemological framework informing the analysis of my data. Specifically, its fine-grained quality suits the questions of the present research, allowing a productive examination of processes of knowledge re-organization. In doing so, it enables me to zoom in on the learning process and analyze the cognitive dynamics of the transitions that occur in knowledge reorganization. As mentioned in Chapter 1, the purpose of this study is to (i) characterize emergent ways of PSTs’ reasoning about sampling variability and (ii) to examine the existence of the contextuality in PSTs’ reasoning processes about sampling variability and, if it existed, to characterize the contextuality. Particularly, this study addresses the following two research questions: (1) What non-normative ideas do PSTs invoke in reasoning about facets of sampling variability? (2) How are PSTs’ reasoning processes about sampling variability sensitive to data-context?

In Chapters 5 and 6 I answer these questions by showing how I developed and used analytical plans informed by Knowledge Analysis (diSessa, Sherin, & Levin, 2016). The steps of the analytical plan are described in Chapters 5 and 6 where they were used. In Chapter 5, I present my analysis of episodes of PSTs’ reasoning related to different aspects of the notion of sampling variability to respond to the first research question. I then used the findings presented in Chapter 5 to adapt and extend Pfannkuch’s (2008) framework concerning ways of thinking about sampling variability. In Chapter 6, I present my analysis of episodes of a PST reasoning related to the contextuality in his reasoning about the notion of sampling variability to respond to the second research question.

---

7 Knowledge Analysis (KA) is a methodological approach that is aligned with the epistemological assumptions of Knowledge in Pieces.
CHAPTER 5: EXTENDING A CONCEPTUAL FRAMEWORK FOR SAMPLING VARIABILITY

Introduction

In this chapter, the first analytic chapter in this dissertation, I present episodes of PSTs’ reasoning related to different statistical aspects of the notion of sampling variability—referred to as facets of sampling variability in this study. I use findings of the data analysis presented in this chapter to respond to the first research question “What non-normative ideas do PSTs invoke in reasoning about facets of sampling variability?” I also use the findings to adapt and extend Pfannkuch’s (2008) framework concerning ways of thinking about sampling variability. The data discussed in this chapter came from the transcripts of the interviews of all seven participating PSTs: Judy, Alisha, Bella, Emma, Liza, Susan, and Tanner (pseudonyms). The interview questions explored sampling variability in three different contextual situations: The Bean Task, the Voting Task, and the Gym Task (see Appendix F).

Recall that each interview began with the Bean Task (see Appendix F) where PSTs were presented with a container of red and white beans (population) in which the number of beans of each color was given. Two scoops of different sizes were used to select samples as shown in Figure 5. The second task during the interview was the Voting Task (growing number of samples). In this task, PSTs were asked to use a pre-designed TinkerPlots’ sampler that simulated collecting samples from California state voters who have been asked whether they would vote for or against Proposition 223. The data-context of the third task during the interview was two given dot plots of sample means calculated from random samples selected from a population of 800 gym members. These two sets of samples were selected by two different students to investigate the question “What is the typical time spent at the gym?” As shown in Figure 6, the upper dot plot has more sampling variability compared to the lower one although
they have the same number of sample means. Note that PSTs were intentionally not told the size of the samples that each student collected.

Interestingly, some of the participating PSTs provided contrasting reasoning across the three situations. In particular, their responses included both normative and non-normative reasoning. However, since the focus of this chapter is on the patterns of reasoning present across all of the PSTs in my sample (as opposed to tracing the development of particular PSTs’ patterns of reasoning across their individual interviews), I took into consideration all the data about how the PSTs reasoned about the tasks in order to capture as wide a variety as possible of ways of thinking about sampling variability. The non-normative reasoning that I observed is important because, in reference to what is currently available in the research literature on student and PST reasoning about sampling variability, it sheds light on some ways of learners’ thinking about sampling variability that had not been documented, as well as elaborating ways of thinking that have been found in the literature. While this analysis focuses on non-normative reasoning patterns, the PSTs in the study provided normative reasoning patterns in the interview for other data contexts (and as such, provides an indication that contextuality of reasoning was a prevalent feature in the data corpus – students shared normative reasoning patterns in some data contexts and non-normative reasoning in others). Although the data collected during the classroom activity was considered when I was conducting this analysis, it didn't find clear non-normative reasoning patterns that could be highlighted in this analysis. In the classroom activity, all of the participating PSTs preferred large samples for making inferences because large samples better represent the population. Recall that the purpose of the classroom activity was to use TinkerPlots sampler to investigate the effect of increasing the size of the samples on sampling variability. In the following section, I describe the method used to analyze the data in this chapter.
Analytical Approach

The data analysis reported in this chapter started with reading each of the seven interview transcripts and adding observational comments (descriptive interpretations) to each line in the transcripts about about (1) behavioral indicators such as what the speaker did as he spoke, (2) affective indicators such as puzzlement or confidence, bolstered by observations related to fluidity of speech (e.g., pauses), (3) what the speaker was referring to, or what appeared on the screen or the board at that moment, and (4) any changes in strategies used by PSTs.

In this layer of analysis, I became interested in describing the non-normative ideas students invoked as they reasoned about the tasks. I read the whole transcript carefully looking for any used non-normative reasoning about sampling variability. I schematized these instances, developing a short description for each, and then used these codes that arose from my initial analysis of the data, to look across the entire corpus of interviews. I collected these observations together in a chart that gives a birds’ eye view of the patterns of non-normative reasoning present in my data. After that, I moved from the static picture of using non-normative reasoning in different data-contexts to how they are related, similar, or different and how frequent the similar non-normative reasoning appeared in the transcript. I went through the entire process of reasoning again and described relationships between the used ideas. Across all interviews, I made analytic memos about the potential influences of the previous instruction or the clues included in the interviewee’s language.

Operationally, I examined the PSTs responses to the tasks and my follow-up questions to understand how they were thinking about the tasks. Within a response to a task, a PST might invoke several smaller ideas that comprised their explanation or response. Thus, my bottom-up analysis of their reasoning started at the level of these smaller units of their explanation up until
the extraction of similar reasoning (patterns) about some statistical ideas. I focused in particular on the frequency of the appearance of similar non-normative patterns of reasoning across subjects and also across data-contexts.

As mentioned above, I use “non-normative reasoning” to indicate reasoning that is either statistically inaccurate or not applicable to the data-context at hand (even if it also involved some statistically accurate ideas within the larger explanation). For example, the following explanation by Tanner (T) was considered an example of non-normative reasoning because he assumed that large samples better represent the population and therefore the outcomes of large samples vary to match the natural variation that exists in any population.

O: Why do you think that increasing the size of the samples here [the Gym Task] will not make it [the range] shrink?

T: Because it’s [large samples] more representative of the data. So, I guess it would grow because you’re asking more or you’re using more people in your sample, so you’re getting more variety in your answers. So, I guess it’s more representative overall because you’re asking more people. So, I guess this would also…the range would also grow for this [larger sample].

Although the above explanation assumes that larger samples are more representative of the population (which is a statistically accurate idea), assuming that the outcomes of the large samples will vary or be more spread out makes this whole explanation non-normative because this last idea is statistically inaccurate. Note that here I refer to the reasoning or the explanation.

---

8In this analysis, I refer to myself as “O” or “I” in the transcription. The transcription conventions used are the following: (a) “[…]” for a break in the speech, typically including a pause, when restart or new direction; (b) “[Italic]” for interpretive and informal commentary, including references to particular displays; (c) pictures of the sampling outcomes or the screens are embedded in the text; and (d) no deletions have been made from the transcript segments provided.
leading to the answer rather than the answer itself (which is also shown in the above excerpt of the transcript). That is, “So, I guess it would grow” was Tanner’s answer for my question about the effect of the sample size on the range, but my analysis included all of the talks that led up to and/or followed answers to specific questions.

**Pfannkuch’s (2008) Framework for Thinking about Sampling Variability**

As mentioned in chapter two, Pfannkuch (2008) developed a framework in which she identified five ways of thinking by which high school students develop their understanding of sampling variability. She used the word “image” to refer to something learners can imagine or conjure up in their minds about five different facets of sampling variability. In the following, I discuss each of these five ways of thinking because they inform the data analysis presented in this chapter. First, “Image of sample size effect” in this framework represents students’ ways of thinking about the effect of the sample size as a facet of sampling variability. This image involves noticing that as the size or the number of the samples increases some value appears clearly in the sampling distribution and therefore information can be concluded about the distribution of the population. Also, according to the framework, a learner should notice the variability among samples with the same size and the variability among samples with different sizes. Dierdorp et al., (2012) used Pfannkuch’s framework and named this image as “sample size and law of large numbers.” Second, “Image of intuitive confidence interval” involves awareness of the effect of the sample size on the location of the expected value in the sampling distribution. In Pfannkuch’s study, three box plots were used to represent sampling outcomes associated with three sample sizes. The goal of using these three box plots was to invoke students’ thinking about the effect of the sample size on the shape of the box plots. Third, “Image of distribution” involves awareness of the shape of the sampling distribution and how it grows to become
symmetric as the number of the selected samples increases. Fourth, “Concept of random process” involves an awareness of the underlying random process of selecting samples and how random samples behave in terms of the data-context. Student thinking also involves an awareness of the sampling method and its effect on the sampling outcomes. Lastly, “Image of the relationship of sample distribution and population distribution” involves an awareness of the overall purpose of selecting samples which is making an inference about some aspect or characteristic of the parent population.

The statistical background of the high school students in Pfannkuch’s study was comparable to the statistical background of the elementary/middle school PSTs in my study and so it was reasonable to expect that the framework would provide a solid foundation for my study. I was informed by Pfannkuch’s (2008) framework in developing the questions included in the interview protocol (see Appendix F) in a broader sense but these questions were specifically chosen to be broad enough to allow the PSTs to discuss a variety of facets of sampling variability. That said, some of the interview questions were specifically designed to follow-up on potentially interesting findings from my Pilot Study. For example, one of the participating PSTs in the Pilot Study, Sophia, had an interesting pattern of reasoning about one facet of sampling variability. That is, even with 100 samples of size 10 selected using TinkerPlots, that clearly show a cluster around some value, Sophia wasn’t convinced that this many small samples might represent the parent population. On the other hand, she was very convinced that one large sample of size 500 well represented the population. Therefore, I added some questions to the interview protocol to provoke the participating PSTs to reason about this facet that I wanted to investigate further.
While I was informed by Pfannkuch’s framework in a general sense in designing my interview tasks, I did not set out *a priori* to code my data using Pfannkuch’s framework. However, as I analyzed my data in the bottom-up way described above, I recognized that the PSTs appeared to be using patterns of reasoning that were not explicitly articulated in Pfannkuch’s framework. It thus became of interest to me to explore in more detail the dialogue between this framework and my data. Accordingly, I revised the first research question to “What non-normative ideas do PSTs invoke in reasoning about facets of sampling variability?”

**Patterns of Reasoning about Sampling Variability Appearing in the Data**

At the point in the semester when I conducted the interviews (week five), the class had studied sampling methods in addition to an introduction to sampling distribution and variability through an activity in which they investigated the length of the words in the Gettysburg Address by selecting words randomly from this address and developing a sampling distribution. The main goal of this activity was to study the notion of randomness as the lack of pattern or predictability in events rather than haphazard. That is, class discussion during this activity highlighted that they could not randomly pick words from the Gettysburg Address as they each have some kind of inherent bias as they chose words. There was some class discussion about sampling variability (mainly concerning the shape of the sampling distribution and the expected value) during the Gettysburg Address activity, but the class had not yet carefully studied measures of center or variability.

Given that one of the central purposes of this dissertation study was to describe the nature of PSTs’ knowledge and reasoning about sampling variability, the data analysis for the study proceeded using bottom-up methods of open coding (as opposed to top-down methods of coding
for predetermined facets of sampling variability). I decided to pay particular attention to patterns of reasoning that were non-normative and used by at least two different PSTs.

I started the data analysis with adding observational comments (descriptive interpretations) to each line in the transcripts about, for example, what the speaker was referring to, or what appeared on the screen or the board at that moment. After that, I used a qualitative data analysis software, Dedoose, to help with finding any non-normative reasoning in the transcript and noticing the frequency of the appearance of similar non-normative reasoning (patterns) across subjects and across data-contexts. In this layer of analysis, I read the whole transcript, which was uploaded to Dedoose, carefully looking for any episodes that might illustrate non-normative reasoning. I pulled out all of these episodes from the transcript and listed them in a table. These episodes of non-normative reasoning fell into four different categories based on the statistical ideas that they entail. These four categories, which are referred to as facets of sampling variability, were: (1) Selecting many small samples then combining them into one large sample vs. finding the sample statistic for each of them separately, (2) Selecting many small samples then combining them vs selecting another large sample with the same number of subjects, (3) Determining the effect of the sample size on the range and the clusters, and (4) Deciding whether more or less sampling variability is desirable. To help frame the following discussion on the results of the interviews, Table 5 presents the four facets of sampling variability along with the non-normative answers corresponds to each of them.

Table 5
Four facets of sampling variability along with the non-normative answers

<table>
<thead>
<tr>
<th>Sampling variability facet</th>
<th>PSTs’ non-normative answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selecting many small samples then combining them into one large sample to find a sample statistic vs finding a sample statistic for each of them separately and noting the distribution.</td>
<td>Selecting many small samples then combining them into one large sample is better than finding the sample statistic for each of them separately.</td>
</tr>
<tr>
<td>Selecting many small samples then combining them vs selecting another large sample with the same number of subjects</td>
<td>Selecting a large sample is better than selecting many small samples then combining them.</td>
</tr>
<tr>
<td>The effect of the sample size on the range and the clusters</td>
<td>The range will grow as the sample size increases.</td>
</tr>
<tr>
<td>More or less sampling variability is desirable</td>
<td>More sampling variability is better for making inferences.</td>
</tr>
</tbody>
</table>

In the following, I discuss in turn the PSTs’ non-normative patterns of reasoning about these four facets. That is, the reasoning provided to justify each of these non-normative answers. I then use these non-normative patterns of reasoning to clarify some of the images already identified in Pfannkuch’s framework, also adding new images.

**First Facet – Selecting Many Small Samples Then Combining Them into One Large Sample vs Finding the Sample Statistic for Each of them Separately**

During the interview, some participating PSTs provided interesting and inconsistent responses to my question about the tradeoff between finding the sample statistic for each of the selected small samples separately and combining them into one large sample. Statistically, these two inferential strategies are equivalent because displaying the sample statistic for many small samples will show a clear cluster around the same expected value. In the following, I discuss
some non-normative reasoning patterns in which PSTs either used or preferred combining small samples into one large sample, then finding one sample statistic for this large sample. First, Bella (B) combined the small samples that she selected in the Bean Task using the small scoop. This was her intuitive strategy when I asked her to use the small scoop instead of the large one. She combined five small samples of beans then found the percentage of the red beans in the new large sample as shown in the following excerpt.

O: Let’s say that it’s important for us to make a prediction about this population and we don’t have any scoop rather than this. We have to use it. Can we come up with some idea, some way by which we can make a good estimation for this?

B: Right. If you fill it up all the way and maybe if you repeat that several times and put them into a pile

O: Can you do this?

B: Sure. So, I do one. Maybe I would do five. Two. Three. Four. And one more. Okay. Then from there, you can count the white versus the red. And just, just by a visual glance, it looks like a pretty good representation of the population.

I argue that describing the resulting large sample as “pretty good representation of the population” in the above excerpt indicates that Bella was thinking about the representativeness of the selected sample foremost and therefore the idea that large samples are always more representative than small samples is the reason behind why she chose to put the small samples together instead of finding the statistic of each one of them separately.

Similar to Bella, Liza also suggested selecting many small samples then combining them into one large sample. She selected 10 small samples in the Bean Task and poured them in 10
separate piles on the table. When I asked her what she was going to do next, she gave the following response:

L: Okay. Then I would like to separate the colors then. Like kind of like how, as I would, I would take this data, I would take all the noes [red beans] and all the yesses [white beans] and count them up and put them on like a chart, like a graph of some sort, like how we do with tinker plots, kind of. And like have it be like yes and no and then like how many people, and do a bar graph, how much, because obviously there are more yesses than there are nos. We would see how much more yesses there are than nos.

While she was working on the Bean Task, Liza told me that she would like to assume that white beans represent the people who voted “yes” and the red beans represent the people who voted “no.” Although Liza has kept her small samples separate, she was talking in the above response about finding the total number (frequency) of beans of each color then presenting the two frequencies using a dot plot. This strategy, however, is equivalent to gathering the small samples then finding the sample statistic in the resulting large sample.

Similar to both Bella and Liza, Alisha preferred combining the small samples together in the Bean Task as shown below.

O: Which method do you trust more? This pile or these separate piles where you think about each of them separately?

A: I vote we do that one [the large sample]

O: Why that one?
A: Because with this one [separate small samples], you’re going to have a whole bunch of different percentages. This, you’re going to have a percentage of white compared to a percentage of red.

Alisha preferred finding the sample statistic for all of the small samples together because that will give a single value (estimation) rather than many values. That is, she preferred combining the small samples (without doing that) because it will be easier for her to deal with a single percentage rather than “a whole bunch of different percentages.”

As mentioned before, most of the participating PSTs provided, in other data-contexts, some normative reasoning patterns that interestingly contrast with the non-normative reasoning discussed in this chapter. However, not one of the participating PSTs provided a normative reasoning associated with the facet of sampling variability discussed in this section namely, the equivalence between combining many small samples to make one large sample then finding a single sample statistic on one hand versus finding the statistic of each one of them separately and then noticing any pattern in the sampling distribution on the other hand. The above three examples of reasoning exhibited a non-normative pattern of reasoning about this facet of sampling variability. Because of the pervasiveness of this idea, I consider this idea as a facet of sampling variability of PSTs’ thinking to be one that needs to be highlighted in Pfannkuch’s (2008) framework.

**The connection of this facet to Pfannkuch’s framework.** The “image of concept of random process” in Pfannkuch’s (2008) framework involves an awareness of the underlying random process of selecting samples and how random samples behave in terms of the data-context. It also involves an awareness of the sampling method and its effect on the sampling outcomes. In Pfannkuch’s discussion of this image, there was no consideration for the
fundamental idea of the equivalence between selecting enough small samples and noticing a pattern on one hand versus combining these small samples into one large sample then calculating the sample statistic for this sample on the other hand. Selecting (or simulating) the selection of multiple small samples is important for understanding the concept of sampling variability because it leads to building the sampling distribution and therefore understanding the underlying logic of the law of the large numbers. One of the fundamental ideas that underpins and validates selecting multiple small samples is the equivalence between selecting enough small samples and noticing a pattern on one hand versus combining these small samples in one large sample then calculating the sample statistic for this sample on the other hand. With that being said, elementary and middle grades PSTs need to be convinced about this equivalence between these two sampling strategies. The non-normative pattern of reasoning associated with this idea highlighted in the above excerpts raises the need for addressing this idea explicitly in Pfannkuch’s (2008) framework for the ways of thinking about sampling variability under the “image of concept of random process.”

**Second Facet – Selecting Many Small Samples Then Combining them vs. Selecting Another Large Sample with the Same Number of Subjects**

During the interview, some participating PSTs provided non-normative reasoning about the tradeoff between selecting many small samples then combining them in one large sample and selecting another large sample. First of all, let us clarify the difference between the facet of sampling variability discussed in the previous section and this facet. In particular, the comparison in the previous facet is between two arrangements of the same set of subjects selected once from the population whereas the comparison in this facet is between two different sets of subjects selected using two different strategies. Statistically, combining many small samples is equivalent to selecting one large sample as long as they have a similar number of
subjects. In the following, I present two categories of PSTs’ non-normative reasoning about this idea.

i. Selecting One Large Sample is More Reliable than Selecting Many Small Samples Then Combining them in One Large Sample

Some PSTs argued that one large sample is more reliable than many small samples and justified their responses differently. Emma was asked in the Bean Task to compare a sample of beans selected using the large scoop and multiple samples selected using the small scoop. Although Emma showed some hesitation when she was answering this question, she ended up saying that one large sample is more representative because large samples are always more representative than small samples.

O: Assume that I changed my rule and allowed you to collect the same number of beans using these two scoops and put them in two piles outside the box. Which pile would you trust more as a sample that represents the population?

E: I think, I think I... trying to think. I like this [E was pointing to the large scoop] method, I think.

O: Why?

E: I don’t know if I’m thinking correctly but like that just, when you’re, that’s a sample of the population and you only have to do one scoop. For this [E was pointing to the small scoop], you had to do like five scoops. Trying to think. __

But I just feel like this is a better... I think it’s a better representation of this population. I don’t, I don’t know how to think about this question.

O: Does that affect the certainty or level of confidence that you have?
E: Yeah. I always trust large samples more than small samples.

To ensure that Emma was aware that the number of subjects in both cases was the same, I questioned: “Even if the total was the same?” She said yes and added that the combined samples are less representative because they will make a large sample that is “made up of small samples.”

Similarly, in the following excerpt, Susan (S) demonstrated her preference for one large sample selected using the large scoop during the Bean Task over many combined small samples selected using the small scoop.

O: Which one of them you trust more [one large sample or five small samples together]?

S: Probably the large scoop because if that’s like in, I don’t want to say equal but like when you’re choosing five times in a row out of the small one, then you’re getting like different sets of data every single time you’re pushing them all together. In the large one, it’s all like one set, just one time that you’re getting it.

Susan clearly preferred the large sample, yet she also seemed to not prefer small samples because their outcomes will be “different” or not consistent. So far, we have seen two examples of PSTs who provided similar but non-normative reasoning about the equivalence of selecting one large sample and selecting many small samples then combining them in one large sample. According to my criterion, these two examples are enough to consider this idea as a facet of sampling variability. However, four more examples of non-normative reasoning about this idea, but with different reasoning, will be seen in the next section.
ii. Selecting Multiple Small Samples Then Combining them is More Reliable than Selecting One Large Sample.

An argument that was observed across some of the interviews was that combining multiple small samples is more representative than selecting one large sample. Different justifications were given for this argument. For example, in the Bean Task, Tanner argued that multiple samples are more reliable than one large sample because each one of these small samples represents a trial. Therefore, many trials are better than one trial.

T: Maybe, right away, I would think the smaller one because you went in five separate times so it’s kind of like doing five separate trials.

Later in the Voting Task, he provided the same answer when reasoning about the issue in another data-context. This time, however, he provided more details and clearly explained his reasoning as shown below.

T: Because for the smaller one, it’s like every little dot has its own like sample attached to it that has its own mean and like all its own data, whereas the one big one just has one mean. So I mean, we’ll see if it’s the same but I would think that this one [many small samples] is more accurate.

Tanner argued that multiple small samples are more reliable than one large sample because each of them will provide an estimation (mean) whereas the larger sample will give one estimation only. In the Voting Task, Emma also said that many small samples will be more representative of the population when combined as shown below.
E: I just feel like there’s, there’s a lot of trials in there, where the 1500 people, that’s one trial, one percentage. I just feel like the more trials we have, the better accuracy there is.

E: I just feel like there are more trials of like, so you have each trial, what’s it called? There are 15 people in each trial and there’s 100, yeah, there’s 100 trials so I just feel like there’s a lot of trials in there, where the 1500 people, that’s one trial, one percentage. I just feel like the more trials we have, the better accuracy there is.

She argued that one large sample represents one trial (regardless of the size) whereas multiple samples represent multiple trials (which she considers as more representative). Like Tanner, Emma added that each sample gives one percentage. Similarly, Susan (S) preferred in the following excerpt—during the Voting Task—many small samples when combined over one large sample just because each of these small samples gives a percentage (an estimation) whereas the large sample gives one estimation only.

O: What do you think now? Which sample or way of sampling do you trust more?

This one or the previous one?

S: With this you can, I guess you only see like one single number, whereas the other one, you get to actually look at what all of the percents, like across the 100 groups were so you can… I mean like visually, you can look at it and trust it more because there’s more of like a representation of like each group having that same percent over and over and over again, compared to this one where you’re just having one group of 65%.
The last case of preferring many small samples over one large sample is Alisha who argued in the following excerpt—during the Voting Task—that many small samples are more reliable because they represent more trials.

O: Which way you trust more?

A: The first way.

O: Why?

A: Because we’re only getting one…

O: It’s not one person. This is one percentage that came out of one sample.

A: so we only really did one sample then

O: Yeah, one large sample.

A: I would rather have a smaller sample and more trials or more runs or collections if that makes sense. We’re still in that range though that I said.

Despite the above examples of the non-normative reasoning about this facet of sampling variability, some of the participating PSTs provided normative reasoning about this facet. Below I give examples of these normative reasoning just for clarification. During the Voting Task, Bella was asked whether she prefers selecting 100 samples each of size 15 or selecting one sample of size 1500 if she makes an inference about Proposition 223. She first wondered about the randomness of selecting these 100 small samples.

O: So instead of selecting these samples [I pointed to the screen where the 100 samples were shown on TinkerPlots], I will select one sample, only one sample of size 1500. Which size do you prefer?
B: I don’t know if I can decide because I don’t know where the samples are coming from.

When I affirmed that these 100 samples are selected randomly without duplicates, she answered my question by arguing that these two options are the same as shown below.

O: Selected randomly

B: Selected randomly, okay. Then I don’t really think it makes much difference if the small ones are selected randomly and the big one is selected randomly. I feel like it doesn’t make much difference.

I tried to challenge her argument by reminding her that each of these 100 samples is small therefore less reliable. She continued to insist that these two choices are the same as long as they will select the same number of subjects.

O: But these [the samples of size 15] are small.

B: Right.

O: Why it doesn’t make difference?

B: Because like I said before, if you’re taking small samples, of these people [California state people], and then you’re asking a totally different set in the next small sample and a totally different set in the next one, eventually, you’re going to have different people that add up to the same number as if you took them all at once randomly. So, they’re both random and they both have the same number.

In the next section, I describe a connection between Pfannkuch’s framework and the facet of sampling variability identified using the above patterns of non-normative reasoning.
The Connection of this Facet with Pfannkuch’s Framework. Similar to the facet of sampling identified in the previous section, the equivalence of selecting multiple small samples then combining them on one hand versus selecting one large sample, on the other hand, is a fundamental idea that underpins and validates selecting many small samples. These two facets are of special importance to beginning learners who tend to intuitively prefer large samples. The “image of concept of random process” in Pfannkuch’s (2008) framework, however, also didn’t consider this fundamental idea of sampling which appeared to be unclear for some PSTs in this study. The above PSTs’ non-normative reasoning about this facet of sampling variability highlights the importance of addressing it explicitly under the “image of concept of random process” in Pfannkuch’s (2008) framework.

Third Facet – The Effect of the Sample Size on the Range and the Clusters

During the interview, some participating PSTs provided non-normative reasoning when they answered my question about the effect of the sample size on the range of the sampling distribution and the clusters of data. Statistically, increasing the size of the samples reduces sampling variability, therefore, decreases the range of the sampling distribution and makes the sample statistics (data values) closer to each other or more compact. However, some participating PSTs in this study weren’t successful in describing this effect of the sample size on the range and cluster in some data-contexts during the interview (although they managed to describe it accurately in other data-contexts). Also, those PSTs provided non-normative patterns of reasoning when they justified their answers. In the following, I discuss some episodes in which PSTs argued that the range of the sampling distribution will either stay the same or increase as the size of the sample increases.
The first episode was when Liza reasoned in the Bean Task that using the large scoop instead of the small one to select the samples would not affect the percentages of the red beans in the selected samples.

O: I’m not going to do it, but I’m just asking you, what if you use this \textit{large scoop} instead of this \textit{small scoop} to make these ten samples? What would happen?

L: I think they would stay, I think they would be around the same.

O: Why?

L: Because you have like a bigger scooper so like you’re going to get more red ones but you’re also going to get more white ones, too. So, like I feel like the ratios of the two are still going to stay the same.

Liza justified her answer by arguing that the ratio of the white to red beans inside the large scoop will stay the same and therefore the percentage of the red beans will stay the same. Thus, the overall sampling outcomes would not change. In response to my follow-up question about the effect of using the large scoop on the range of the sampling distribution, Liza first said that the range would grow, but then she immediately retreated and said it would be the same. I conjecture that she remembered her answer in the previous question in which she said that the sampling outcomes would not change therefore the range, which is dependent on these sampling outcomes, would also not change.

O: Okay. If we conducted this experiment using this scoop, can you expect the range?

Can you say something about the range?

L: The range would be larger.
O: Larger?

L: Yes. Because you kind of have more red beans. But then you’re also going to have more white beans, too… I don’t know. I feel like it’s going to be like similar. I know it’s probably not right, but I feel like it’s going to be similar.

O: I see.

L: Because like, like this might, like this might be 25% [She pointed to a small sample in which 25% of the beans are red] right here but if I count all these [She pointed to the large sample selected using the large scoop], this might also be 25%, just like, you know, it’s just, but this is like a simplified version of this could be kind of thing.

O: So, in terms of the range, you don’t expect anything to happen with this scoop?

L: I mean, I think… No, I think the range will be about the same.

Although Liza seemed to hesitate in her response, when I asked her to summarize her answer about the range, she confirmed that the range would stay the same. Similar to Liza, Judy thought that the percentage of the red beans would not change in the large scoop.

O: Okay. What if I repeat this experiment using this scoop [the large scoop]? I don’t want to do it.

J: Okay

O: I’m just asking. What do you expect the outcomes to be?

J: Probably pretty similar.
O: Why?

J: Like the numbers would be different but like I think the percentages would be the same because like the number didn’t really change. Like the numbers of each color didn’t really change.

Judy, however, provided normative reasoning about the effect of the sample size on the range when she answered similar questions in the two other data-contexts. The above two non-normative reasoning provided by Liza and Judy are enough (according to the suggested criterion) to consider the effect of the sample size on the range as a facet of sampling variability that is worth highlighting. However, one more example of non-normative reasoning about this facet was provided by Tanner, but with a different argument. He argued in the Gym Task that the range will grow as the size of the samples increases. He justified this by saying that large samples better represent the population and therefore their outcomes vary to match the natural variation that exists in any population.

T: Because it’s more representative of the data. So, I guess it would grow because you’re asking more or you’re using more people in your sample, so you’re getting more variety in your answers. So I guess it’s more representative overall because you’re asking more people. So I guess this would also, the range would also grow for this.

Despite the above examples of the non-normative reasoning about this facet of sampling variability, some of the participating PSTs provided normative reasoning about this facet. Below I give examples of these normative reasoning just for clarification. As opposed to her non-normative reasoning in the Voting Task, Judy provided normative reasoning in the Gym Task.
about the effect of the sample size on the range of the sampling distribution. I first asked her to describe the difference between the given sampling distribution in order to let her notice that one of them has a wider range than the other. The next step was to ask her about the reason that might have caused this difference although each of the two students who investigated the question “What is the typical time spent at the gym?” selected the same number of samples. Judy easily noticed the differences between the sampling distributions including the range.

O: Could you please describe the difference between their [the two students who selected two sets of samples to investigate] sampling outcomes?

J: The top one [the top sampling distribution in Figure 6] is a lot more spread out.

So, it’s a bigger range.

When I asked her about what might have caused this difference, Judy didn’t give any reason. I judiciously asked her if the size of the sample might have caused this difference then she said yes as shown below.

O: Can you think about anything that might have caused this difference [between the two sampling distributions]?

J: No

O: What if I tell you that each of them, each of the students, used a different size for his samples? Does that matter?

J: Oh, yeah

When I asked her to justify why the size matters, she tried to say that as the size of the samples increases, the means of the samples will be closer to some specific value that she called “the answer.”
O: Why?

J: Well, because if you use more [more subjects in the samples], I think each mean [the sample mean] would be like closer to like the, like would be closer to the… I don’t know. Something. To the answer.

I argue that the above response reflects an awareness of the effect of the sample size on the shape of the sampling distribution because increasing the size of the sample, according to her argument, will result in sampling outcomes that are closer to some specific value. Therefore, the overall shape of the sampling distribution will be more compact with a smaller range. However, Judy mentioned the effect of the sample size on the range explicitly in the next response as shown below.

O: So, which one of them do you think used the larger size?

J: The bottom one [the sampling distribution with a smaller range in Figure 6]

O: The bottom one used a larger size?

J: Yeah, because it's a lot smaller, like range of answers.

**The Connection of this Facet to Pfannkuch’s Framework.** One of the ways of thinking about sampling variability identified in Pfannkuch’s (2008) framework is the “image of intuitive confidence interval.” This involves creating images of sample statistics intervals for different sample sizes. Although this facet highlights the effect of the sample size on the confidence intervals, there was no clear mention in the description of this facet for the exact effect of the sample size on the range of the sampling distribution. It might be because Pfannkuch’s framework was initially intended to be used with high school students who are expected to use
confidence intervals in their reasoning. The above non-normative reasoning by each of Liza, Judy, and Tanner, however, highlights the need for explicitly addressing the effect of the sample size on the range of the sampling distribution in this framework if we use it to accurately describe elementary and middle grades PSTs' ways of thinking about sampling variability. All of the participating PSTs were successful in describing the effect of increasing the number of samples on the range. That is, they all argued that the range might grow as the number of the samples increases which is an accurate answer statistically. Those PSTs, however, might give non-normative reasoning about the effect of increasing the number of the samples on the range in different data-contexts, therefore, it might be the next research step. Any future research that catches PSTs' non-normative reasoning about the effect of the number of the samples on the range might easily extend this facet.

**Fourth Facet – More or Less Sampling Variability is Desirable**

Statistically, less sampling variability reflects a better sampling method; therefore, the sampling distribution is more reliable for making inferences about the population. Although this idea seems intuitive to PSTs, the data analyzed in this study revealed that some PSTs might have an emerging understanding of the overall purpose of the sampling distribution and the relationship between its shape and the trustworthiness of the drawn inferences. That is, some of the participating PSTs preferred, in some data-contexts, scattered sampling distributions to make inferences and showed non-normative patterns of reasoning as they justified their answers. In the following, I present some of the PSTs' responses in which they preferred scattered sampling distributions to make inferences.
During the Gym Task, Tanner argued that he trusted the sampling distribution with the wider range more if he wanted to make an inference about the population as shown in the following excerpt.

T: Because the data is more like, I don’t know. Maybe I would do the first one [the sampling distribution with the wider range] actually because like I was going, like I was saying before, the range makes me like verify, I think like there was like a good method of sampling done because there is such a range of means, rather than the second one is kind of just all pushed together. So, I guess the first one, even though it might not visually like helps me make the assumption of what the mean would be. But I feel like it is, like since it is such a variety, you can kind of find the mean of all of those and it will be your answer even though those like outliers will affect them. I feel like that’s kind of what they’re supposed to do so I mean, like if that many people in the sample said 34, then it should affect it like should affect your answer in finding the mean, I think.

Tanner justified his answer by saying that a sampling distribution with a wider range indicates a better sampling method, therefore, it better represents the parent population. Susan also preferred, in the Gym Task, the sampling distribution with the wider range because, similar to Tanner, she thought it better represented the diversity of the population.

O: Which of the dot plots do you trust more if you would like to make a prediction or a conclusion about the population?
S: I mean, one with the wider range so you can see how the whole population between that whole range is represented. There’s, so you have, I don’t know, a wider range to look at.

The above two non-normative reasoning provided by Tanner and Susan are enough, according to the suggested criterion, to consider the relationship between the shape of the sampling distribution (especially the range) and the trustworthiness of the drawn inferences as a facet of sampling variability that is worth highlighting. However, one more example of a non-normative reasoning, but with a different argument, about this facet was provided by Emma. In the Gym Task, Emma wasn’t firm in preferring a sampling distribution with a smaller range for making inferences. Although she wanted the sampling outcomes to be different, she also wanted the data to be spread out a little bit so that she could read them easily.

O: Would you like the sampling outcomes, the values that you get out of the sampling, to be similar or different?

E: I would want them to be different, but I would want there to be some similarity or be able to tell the average. But I still want them to be different so it’s not like hard to read the history. If that makes sense.

I claim that Emma wasn't aware of the relationship between the trustworthiness of the inference and the shape of the sampling distribution because if she was aware of this relationship, then she would not consider the readability of the sampling distribution which has nothing to do with the fundamental purpose of sampling distributions and making inferences.

Beside the above examples of the non-normative reasoning about this facet of sampling variability, some of the participating PSTs provided normative reasoning about this facet. Below
I present examples of these normative reasoning just for comparison and clarification. In the Bean Task, Liza wanted the sampling outcomes to be similar, less sampling variability essentially, if she makes an inference about the population.

O: So, if you make sampling like this, would you like your outcomes to be similar or different? If you would like to make a prediction about the population?

L: Similar

O: Why similar?

L: Because then you can predict patterns for the future. If it’s all, if it’s all spread out and varied, then you won’t be able to make predictions unless we do more sampling.

She justified the above answer by arguing that similar outcomes will indicate “future” pattern. I consider these reasoning as normative because each of them is both statistically accurate and applicable in the data-contexts at hand.

**The Connection of this Facet to Pfannkuch’s Framework.** The primary goal of presenting the sampling outcomes in a sampling distribution is to aid in the process of making accurate inferences about the parent population. Theses accurate inferences require sampling distributions with small ranges or clear clusters. Therefore, there is an inverse relationship between the range and the clusters of the sampling distribution on the one hand and the trustworthiness of the sampling distribution on the other hand. This inverse relationship, however, seems unclear to some of the participating PSTs. Some of them preferred sampling distributions with large ranges to make inferences in some of the interview data-contexts. Pfannkuch’s (2008) framework, however, doesn’t consider student’s thinking about this
important facet of sampling variability. Given that the present study is focused on elementary and middle grades PSTs, I consider PSTs thinking about the foundational idea of this inverse relationship as necessary to be addressed as a new facet of sampling variability.

**Summary of the Non-Normative Patterns of Reasoning**

To visually summarize the above discussion, Table 6 presents the non-normative answers about the identified four facets of sampling variability across the three different data-contexts. This table also shows the names of the participating PSTs who provided each of the non-normative answer in each data-context.

Table 6
*Summary of PSTs answers about different facets of sampling variability*

<table>
<thead>
<tr>
<th>Sampling variability facet</th>
<th>PSTs’ non-normative answers</th>
<th>Bean Task</th>
<th>Prop. 223 Task</th>
<th>Gym Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selecting many small samples then combining them into one large sample to find a sample statistic vs finding a sample statistic for each of them separately and noting the distribution.</td>
<td>Selecting many small samples then combining them into one large sample is better than finding the sample statistic for each of them separately</td>
<td>Bella</td>
<td>Alisha</td>
<td>Liza</td>
</tr>
<tr>
<td>Selecting many small samples then combining them vs selecting another large sample with the same number of subjects</td>
<td>Selecting a large sample is better than selecting many small samples then combining them.</td>
<td>Emma</td>
<td>Susan</td>
<td></td>
</tr>
<tr>
<td>The effect of the sample size on the range and the clusters</td>
<td>Selecting many small samples then combining them is better than selecting another large sample with the same number of subjects</td>
<td>Tanner</td>
<td>Tanner</td>
<td>Emma</td>
</tr>
<tr>
<td>More or less sampling variability is desirable</td>
<td>The range will grow as the sample size increases</td>
<td>Tanner</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>The range will stay the same as the sample size increases</td>
<td>Liza</td>
<td>Judy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>More sampling variability is better for making inferences</td>
<td>Tanner</td>
<td>Susan</td>
<td></td>
</tr>
</tbody>
</table>
Table 7 presents the non-normative reasoning that PSTs used to justify their ideas about the four facets of sampling variability across the three different data-contexts. This table also shows the names of the participating PSTs who provided each of the non-normative reasoning in each data-context.

Table 7

<table>
<thead>
<tr>
<th>PSTs' non-normative answer</th>
<th>PSTs' non-normative reasoning</th>
<th>Bean Task</th>
<th>Prop. 223 Task</th>
<th>Gym Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selecting many small samples then combining them into one large sample is better than finding the sample statistic for each of them separately</td>
<td>Large samples are always more representative than small samples</td>
<td>Bella</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Her intuitive strategy when I asked her to use the small scoop instead of the large one was to select many small samples using the small scoop then combine them in one large sample.</td>
<td>Liza</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Combining the small samples is better because it will be easier for to deal with a single percentage rather than “a whole bunch of different percentages.”</td>
<td>Alisha</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selecting a large sample is better than selecting many small samples then combining them.</td>
<td>One large sample is more representative because large samples are always more representative than small samples</td>
<td>Emma</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>One large sample is better because the outcomes of the small samples will be different or not consistent.</td>
<td>Susan</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 7–Continued

<table>
<thead>
<tr>
<th>Selecting many small samples then combining them is better than selecting another large sample with the same number of subjects</th>
<th>Multiple samples are more reliable than one large sample because each one of these small samples represents a trial. Therefore, many trials are better than one trial.</th>
<th>Tanner Emma</th>
<th>Tanner Alisha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Many small samples when combined are more reliable than one large sample just because each of these small samples gives a percentage (an estimation) whereas the large sample gives one estimation only.</td>
<td></td>
<td></td>
<td>Susan</td>
</tr>
<tr>
<td>The range will grow as the sample size increases</td>
<td>Large samples better represent the population and therefore their outcomes vary to match the natural variation that exists in any population.</td>
<td></td>
<td>Tanner</td>
</tr>
<tr>
<td>The range will stay the same as the sample size increases</td>
<td>The ratio of the white to red beans inside the large scoop will stay the same and therefore the percentage of the red beans will stay the same. Thus, the range of the sampling distribution would not change.</td>
<td>Liza Judy</td>
<td></td>
</tr>
<tr>
<td>More sampling variability is better for making inferences</td>
<td>Sampling distribution with a wider range indicates a better sampling method, therefore, it better represents the parent population.</td>
<td></td>
<td>Tanner Susan</td>
</tr>
<tr>
<td>Spread out sampling outcomes are preferable because we can read them easily.</td>
<td></td>
<td></td>
<td>Emma</td>
</tr>
</tbody>
</table>

### Chapter Summary

Data presented in this chapter has shown that some of the participating PSTs have an emerging understanding of some foundational ideas related to sampling variability. Namely, they
provided non-normative patterns of reasoning when they answered questions about: (1) Selecting many small samples then combining them into one large sample vs finding the sample statistic for each of them separately, (2) Selecting many small samples then combining them vs selecting another large sample with the same number of subjects, (3) Determining the effect of the sample size on the range and the clusters, and (4) Deciding whether more or less sampling variability is desirable. These four facets of sampling variability identified in this data analysis provided an answer to the first research question of this study which is, “what non-normative ideas do PSTs invoke in reasoning about facets of sampling variability?”

What helped with finding all of these non-normative patterns of reasoning in PSTs' responses was interviewing the participating PSTs early in the semester before they sufficiently study sampling variability in their probability and statistics content class. The use of three different data-contexts during the task-based clinical interview has also helped with revealing these non-normative patterns of reasoning. Most of the PSTs provided normative reasoning about these four facets of sampling variability in some data-contexts and non-normative reasoning in other data-contexts. The patterns of non-normative reasoning presented suggest some adaptations and clarifications to Pfannkuch’s (2008) framework for the ways of thinking about sampling variability. Those will be discussed more in the overall conclusions in Chapter 7.

Looking Ahead to Chapter 6

Previous studies that investigated PSTs reasoning about sampling variability have not shed light on the potential contextual nature of PSTs' reasoning about sampling variability. Based on the findings of Wagner (2006) and the principles of the coordination classes model (diSessa & Sherin, 1998; diSessa & Wagner, 2005; diSessa, Sherin, & Levin, 2016), I hypothesized that PSTs' reasoning about sampling variability might be highly contextual. Thus, a
clear case in which a PST exhibits a high degree of contextuality in their reasoning about sampling variability is of particular interest for my study.

Studying the phenomenon of contextuality in PSTs’ reasoning is not an easy research task because what participants do as they compare data-contexts often happens in a fairly rapid way that it’s hard to observe. That is, studying the phenomenon of contextuality requires very rich data that clearly exhibits details of thinking. To do so, I tried to cast a wide net across subjects and data-contexts in order to raise the possibility of choosing a PST who is able to provide rich and useful data for my analysis. Among the seven participating PSTs, a goal in my mind as I read the interview transcripts was to identify a PST who was particularly vocal and articulate throughout the interview, and who showed signs of active engagement in the tasks through their willingness to wrestle with the tasks and try to justify their answers. In the end, one such case rose to the top, Tanner, whose case I analyze in detail in the next chapter. In the next chapter, I used the facets of sampling variability described in this chapter to provide a fine-grained analysis of the contextuality that appeared in Tanner’s reasoning about sampling variability and provide an answer to the second research question “How are PSTs’ reasoning processes about sampling variability sensitive to data-context?”
CHAPTER 6: CONTEXTUALITY IN THE CASE OF TANNER

The subject described in this chapter, Tanner, is one of the seven PSTs who I interviewed and analyzed their data in Chapter 5. Tanner’s major is elementary teacher education and he had not taken any statistics courses before. The data discussed in this chapter mainly came from videos of both whole- and small-group discussions during one session of the statistics content course during the fall 2018 semester, video recordings and written responses during task-based clinical interview and from the online questionnaire which aimed at providing some background data about the participating PSTs and their current knowledge about sampling variability (see Appendix D). Other data were drawn from the screen recording of the laptop during the interviews and the video recorded one class activity. The interview (see Appendix F) had three tasks, each of which entailed similar reasoning about sampling variability and each in a different data-context. In the following section, I discuss the analytical framework that I designed and used to analyse the data in this chapter.

Analytical Framework Used in this Chapter

To answer the second research question—How are PSTs’ reasoning processes about sampling variability sensitive to data-context? —in this chapter, I report how I developed and used four analytical strands informed by a methodological approach called Knowledge Analysis (diSessa, Sherin, & Levin, 2016). This methodological approach is aligned with the epistemological assumptions of Knowledge in Pieces.

Analytical Strand One – Choosing a PST to be the Case of the Study

I started the data analysis in this chapter by trying to choose a PST who was able to provide rich and useful data for my analysis. Among the cases of contextuality documented in the first part of my analysis, I considered if there was a case that offered (1) multiple instances of
contextuality, (2) particularly good window into thinking processes because of willingness to wrestle with the tasks and try to justify his/her answers. That is, I sought a PST who was particularly vocal and articulate throughout the interview, and who showed signs of active engagement in the tasks through their willingness to wrestle with the tasks and try to justify their answers. Also, I sought a PST who experienced periods of noticeable struggle or uncertainty before successfully recognizing similar tasks as entailing the same facet of sampling variability.

In the end, I chose the case of Tanner to analyze with respect to the in-the-moment contextuality of his reasoning and the dynamics of his reasoning about sampling variability.

**Analytical Strand Two – Finding instances of contextuality:** In this layer of analysis, I read the whole transcript of Tanner’s interview and class activity carefully again looking for any episodes that might illustrate the phenomenon of using different reasoning about the same statistical ideas across different data-contexts. This analytical strand resulted in identifying consequential contextuality episodes from the transcript that are related to some facets of the notion of sampling variability described in previous chapter.

**Analytical Strand Three – Describing contextuality:** In this strand of analysis, I described in detail the sequence of the reasoning that led to each instance of contextuality and described the changes in the reasoning process across the different data-contexts. I also highlighted any potential influences of the previous instruction or the clues included in the instructor or interviewee’s language.

**Analytical Strand Four – Characterizing the phenomenon using the theoretical machinery of Knowledge in Pieces:** By looking in detail at the subjects’ words, I identified some of the constructs associated with the KiP theoretical framework (concept projection, span, alignment, etc.) then used these constructs to characterize in fine-grained detail the phenomenon
using the theoretical machinery of Knowledge in Pieces. I used snapshots of the processes of reasoning about sampling variability to clarify the constructs and the changes in reasoning used by the PSTs.

The organization of the content of this chapter is informed by the strands of the above analytical framework. That is, in the following section I have selected episodes of contextualities from the transcript and have described it (analytical strands two and three). Following this description, I explored the phenomenon of contextuality using the theoretical constructs developed within the Knowledge in Pieces perspective (analytical strand four). Thus, the reader will notice that the analysis is reported in two layers. The first time I present the data from the case of Tanner, I just explain the case and give some illustrative examples of how his reasoning was contextually shaped. I then present a deeper, second layer of analysis that connects that analysis of the data with specific constructs from the theoretical framework.

**Finding and Describing Instances of Contextuality (Analytical Strands Two and Three)**

In this section, I describe in detail two of the clear contextuality examples that the subject exhibited as he was reasoning about the notion of sampling variability in different data-contexts. In the data analysis, I describe the statistical reasoning used by the subject, referring to Pfannkuch’s framework discussed in the previous chapter, in two different contextuality examples.

**Contextuality in Tanner’s Reasoning: Example One**

This section presents an example from the transcript that exhibits how Tanner has given different answers for similar questions about some facet of the notion of sampling variability and used different reasoning to justify these answers across two different data-contexts— the Bean
and the Gym Tasks. This example is related to “image of intuitive confidence interval” from Pfannkuch’s (2008) framework by addressing the effect of the sample size on the range of the sampling distribution. Recall that all interview tasks are in Appendix F.

**Episode One (Reasoning about the Bean Task).** The interview began with a task in which Tanner was presented with a container of red and white beans (population) in which the number of beans of each color was given. Two scoops of different sizes were used to select samples as shown in Figure 5. Tanner was asked to compare between small and large samples with respect to the percentage of the red beans that they contained:

O: What percentage of the red beans do you expect to get in each of the scoops?

T: I guess maybe you’d get 20% in the big one but not the small one because the bigger the sample you take, the more accurate it is to the population.

In this data-context, Tanner showed an understanding of the idea that large samples entail statistical characteristics (mean, proportion, etc.) that are closer to the intended population parameters as compared to small samples. To investigate his understanding of the idea that sampling variability will decrease as the sample size increases, Tanner was asked to select 10 samples using the small scoop and make an inference about the population based on these samples. The percentages that he got in these samples were: 23%, 14%, 33%, 0%, 23%, 26%, 27%, 36, 9%, and 0% as shown in Figure 9.
After that, he was asked about the possible effect of using the larger scoop instead of the small one to select these 10 samples as shown in the following dialogue:

O: Assume we would like to use this [large] scoop instead of this [small scoop] and do the same thing [selecting 10 samples]. We will not do it again. Just assume. What will happen? What would you expect to get using this large scoop?

T: I guess I would expect not as much like variability between.

So, like I probably wouldn’t expect to have 0%’s just because it’s [the larger scoop] bigger so it [the larger scoop] gets more of the population, a bigger sample.

So I would expect it [the percent of the red beans in the larger scoop] to be closer to like the middle of what we found there.

In this data-context, Tanner showed an understanding of the idea that sampling variability will decrease as the sample size increases. I argue that what made Tanner claim in this data-context that the sampling variability will be less with the use of the larger scoop was noticing that one of the samples that he collected using the small scoop had no red beans – something that would be much less likely to occur with the use of the larger scoop. In other words, it would be
extremely unlikely to randomly scoop up the beans in the large scoop and not get at least one red bean. To know why he used the term “variability” in particular, and what did he mean by this term, I asked the following:

O: What do you mean by variability?

T: Like this [the data values on the sampling distribution] ranges all the way from 0 to 36%. I would expect it [the data values] to be like not as much of a range in between there [0 to 36%].

That he mentioned the range in his last answer provides additional evidence that Tanner was aware of the effect of increasing the size of the samples on the sampling outcomes in this data-context. That is, increasing the size of the samples will decrease the range in some way. I think this might be something he noticed in class when he used TinkerPlots to investigate the effect of the size of the samples on the sampling variability, but we can’t tell based on this short answer how solid was his understanding to this notion. However, Tanner showed a different understanding of sampling variability in a later data-context (the Gym Task) during the interview.

**Episode Two (Gym Task).** The data-context of the third task during the interview was two given dot plots of sample means calculated from random samples selected from a population of 800 gym members. These two sets of samples were selected by two different students to investigate the question “What is the typical time spent at the gym?” As shown in Figure 6, the upper dot plot has more sampling variability compared to the lower one although they have the same number of sample means. Note that he was intentionally not told the size of the samples that each student collected. Tanner was asked here about what might have caused this difference between the two dot plots. Based on his answer in the Bean Task, I was expecting him to
mention the effect of the size of the samples as one of the potential reasons behind this difference between the two dot plots. It was assumed in the Gym Task that two students have created the two given sampling distributions and each of them used a different sample size.

O: What do you think might have caused the difference between these two sets of data?

T: I think it could be the method of sampling, like how they asked or yeah, like how they asked people. If one of them just like asked all their friends or if they did it like a different way, like asking every five people or something like that. Which like finding more of an accurate way to represent which is kind of what I was saying with the first one [the upper sampling distribution in Figure 6], that kind of seems more like they did more of a representation of the whole population rather than the second one since it’s more pushed together. Seems like it was more like the same type of people asked.

This answer gives evidence that Tanner’s understanding of the notion that the size of the sample affects the shape of the sampling variability is rudimentary because he argued that the sampling distribution with the wider range represents the large samples. He also added that wider sampling distribution represents a better sampling method because the sampling outcomes show some variability that matches the natural variability exists in the parent population. However, both of these arguments are statistically inaccurate because increasing the size of the samples decreases the range of the sampling distribution and makes the sampling outcomes more compact. While there is some evidence that he may have some understanding that increasing the size of the samples decreases the spread of the sampling distribution as shown previously in the Bean Task, his understanding seemed to be limited to a specific data-context, that of the Bean Task.
After his response above, I asked him to reconsider his reasoning about this task, drawing his attention to the fact that the size of the samples is not the same in the two cases. When I asked him whether that mattered or made any difference, this cued him to start thinking about the influence of the size of the samples on the shape of the distribution. Surprisingly, he now came up with an answer that wasn’t aligned with his previous answer to a similar question in the Gym Task.

O: What if I tell you that each guy used a different sample size? Do you think this might also cause any difference, or does that matter?

T: Yes, I think it does matter because like if the second person sampled less, then that’s why all their stuff may be close together because the amount of people sampled is like more represent… If you have more, a bigger sample, you represent more of the population. So if you sample like ten people and the other person does 50, the other person is going to have a more accurate representation of the population. So it definitely does affect it.

During the interview, when I heard Tanner’s reasoning described above, I wasn’t sure if he meant that the upper dot plot has more samples or the same number of samples as the bottom dot plot but with larger sample sizes. Thus, I asked the following to clarify that both dot plots represent the same number of samples.

O: Which one of them do you think used the larger size in his sample?

T: I think the top one is the larger size.

O: He used, he collected 20 samples [I was pointing to the upper sampling distribution in Figure 6], remember, right? The same.
T: Yeah, more people in each sample.

It thus seemed that Tanner was aware that both dot plots represent the same number of samples and he was talking about the size of the samples rather than their number. With respect to the idea of changing his answer and reasoning, Tanner argued in this response that the upper dot plot, which is more spread out, represents the samples with the larger size. That is, he invoked the idea that larger samples result in more sampling variability. This reasoning, however, contradicts his previous answer in the bean data-context regarding the effect of using the larger scoop on the sampling variability. Not only were his responses in these data-contexts seemingly contradictory, but also the reasoning that he used to justify each of them differed. For example, in his reasoning about task one, he argued that it will be very unlikely to get outliers (a whole sample of white or red beans only) using the larger scoop. He also added that larger scoops (samples) will reduce the range and make the sampling outcomes more clustered. However, his reasoning in the second differed:

O: I feel that you are not confident about this or… still thinking about it?

T: I think, I mean, I am confident in I think saying that the top one used more people and like got more data altogether, but I’m not confident in my answer that the top one is better. But I guess if he did collect more samples, then I would go with the top one or used more people in his samples, then I would be confident in saying the top one is correct or closer to a representative to the population.

In the above segments, Tanner reaffirms his suspicion that “larger samples will result in more sampling variability.” In the following, I reminded Tanner of his previous answer in the Bean Task and pressed him to compare it with his recent answer hoping to understand more about the reason for the inconsistency in his reasoning.
O: Do you remember your answer when I asked you about this set of data, these samples?

[I pointed to the ten beans samples selected using the small scoop. See Figure 9]. I asked you what would happen if I used this scoop [the large scoop] instead of this [the small scoop].

T: No.

O: I think you said that this scoop will give me less variability.

T: Okay. I still agree with that. Because I mean, I guess, […] I think you might have a bigger range if you do ten big samples though because you […] I don’t know. Okay.

So I still think that it’s better to use the big one because you’re getting more of the population. You’re sampling more of the population which I think is always better because you’re getting closer to the most representative of the population.

The above answer shows that Tanner had an understanding that large samples are always better than small samples, but he also thinks that bigger samples increase the range of the sampling outcomes (not correct statistically). However, even when made aware of the discrepancy, it seemed that he couldn’t reconcile these two ideas and kept saying “I don’t know.” In the following, I asked him about the range specifically hoping that he would give a justification for the increase of the range as the size of the samples increases.

O: How about the range? [still talking about the Bean Task. We didn’t go back to the Gym Task yet].

---

9 Indicates a pause in the speech for at least four seconds
T: I guess the range could grow [...] Well, actually, no, I think it will shrink because there’s less of a chance of you getting 0% of red beans with that cup [the large scoop] than there is with that one [the small scoop]. But then there’s also a bigger chance of you getting more than like four beans or 36% of red beans. So, I could see it going either way. I guess it [the range] would shrink just because… but that’s just because I don’t think you’d have 0%, or you’d have less 0%.

In the above, Tanner got confused because, on one hand, he reasoned that large samples will never have 0% of the red beans thus implying that the lower bound of the range will increase. On the other hand, he thinks that large samples will have a greater chance of getting more red beans in them and therefore a bigger percentage of the red beans.

What Tanner missed here is that the larger samples will also contain a larger number of the white beans, which will affect the percentage of red beans. It wasn’t clear why he finally concluded that the range would shrink although the upper bound will increase. Perhaps he thought that the increase of the lower bound exceeded the increase of the upper bound and therefore the range would decrease.

In the next question, I directed him to consider the gym member tasks again to see if he could reconcile his reasoning in the bean data-context with his reasoning about the same issue as in the gym task.

O: Why do you think that increasing the size of the samples here [the Gym Task] will not make it [the range] shrink?

T: Interesting. Because it’s more representative of the data. So, I guess it would grow because you’re asking more or you’re using more people in your sample, so you’re
getting more variety in your answers. So, I guess it’s more representative overall because you’re asking more people. So, I guess this would also, the range would also grow for this.

In the above response, Tanner didn’t make any connection with his answer in the Bean Task. It seems that he didn’t notice that these two questions entail the same facet of the notion of sampling variability i.e., the effect of the sample size on the range of the sampling distribution. In the following segment, I told him that you will never get a 0% if you use a large sample in the Gym Task just like the beans’ data-context.

O: But you will never get a 0% like this [I pointed to the sample that has 0% red beans].

It’s hard.

T: Right, but I think you would get bigger than 36 so just because you’re getting less of this [Tanner pointed to the small sample with 0% red beans], I think you’d still get more of these [Tanner pointed to the sample with the highest percentage of the red beans; 36%].

In the above response, Tanner continued to insist that increasing the size of the samples in the gym data-context would increase the range. He tried to justify this using an argument that differed from the argument he used to reason about the Bean Task. As opposed to his reasoning in the Bean Task in the data-context of the gym members he argued that the upper bound of the range will increase more than the increase of the lower bound therefore the range will become wider. He was confident that increasing the size of the samples would increase the lower bound of the range because he believed he had physical proof for this notion (the sample of 0% red beans). However, he remained unsure about the effect of increasing the size of the samples on the upper bound. In the Bean Task, he said it would increase but in the gym data-context, he said
it would decrease. These contrasting responses reveal that Tanner has an emerging understanding of the idea that increasing the size of the samples will decrease the range by increasing the lower bound and decreasing the upper bound.

Next, I tried to challenge his above answer regarding the range by reminding him that large samples will contain more white beans, too. That is, the percentage of the red beans might not increase and therefore the upper bound of the range will not increase.

O: And if you get more of it [red beans in the large scoop], you’ll also get more white [beans in the large scoop]. The ratio, thinking about the ratio, the percentage would also, would also…

T: Yeah, I think my original point was that I would get more like similar, like instead of the range being from 0 to 36, I’d get more like similar, like a cluster, I guess you could say, of results with just a few outliers, rather than having this big of a range. So now I’m going back to my original answer to that. It [the range] would be smaller, I think.

Here, Tanner went back to his first answer in the bean task in which he said that the sampling variability “would be smaller” as the size of the samples increases. He tried to justify the decrease of the sampling variability by arguing that the cluster would grow because more similar data values will be added. Therefore, this is what he means by decreasing the sampling variability.

Contextuality in Tanner’s Reasoning: Example Two

This section presents another example from the transcript that exhibits how Tanner gave different answers to similar questions about some facet of the notion of sampling variability and
used different reasoning to justify the answers across two different data-contexts— the Beans and the Gym Tasks. This example is related to the inverse relationship between the shape of the sampling distribution and the trustworthiness of the inferences to be drawn about the population which was addressed in Chapter 5.

**Episode One (During the Bean Task).** After the Bean Task, I pointed to the sampling distribution that we developed using the bean samples, *Figure 10*, and asked Tanner whether he preferred the sampling outcomes to be similar or different if he wanted to make an inference about the parent population.

*Figure 10. Recreation of the sampling distribution created during the Bean Task*

Tanner pointed to a sampling distribution, *Figure 11*, from the Voting Task, the second task during the interview, to support his answer as shown below. This sampling distribution was created using *TinkerPlots* as we were working on the Voting Task.

*Figure 11. Screenshot of the sampling distribution created during the Voting Task using TinkerPlots*
O: If you make a conclusion about the population based on the outcomes like this or these outcomes [I pointed to the board where the bean's sampling distribution was drawn, \( \text{Figure 10} \)], would you like the outcomes to be similar or different?

T: I guess similar so that you can see, like, without these being similar, you wouldn’t have that cluster, so you wouldn’t be able to like see exactly where it’s going to be. But also without the differences, you wouldn’t be able to really tell how like extreme it is. Like in this one [T pointed to the screen of the laptop which showed Figure 11], there’s like 55% in between [Tanner pointed to the lower and upper bounds of the range then estimated the range which is \( 94\% - 40\% \approx 55\% \)] and I think that’s a pretty extreme like difference. So, I think mostly, like having the cluster in the middle is more beneficial than having the range, if that makes sense.

The first part of this response exhibits that Tanner would like the sampling variability to be small in order to see clusters that help with making predictions. He continued by arguing that small range is also helpful but in the case of the sampling distribution in \( \text{Figure 11} \), the range, \( 94\% - 40\% \approx 55\% \), is large and therefore not helpful for making predictions.

\textbf{Episode Two (During the Gym Task).} Later during the Gym Task, I asked Tanner again about the inverse relationship between sampling variability and the confidence in the inferences to be drawn about the parent population based on the sampling distribution. Before that, I asked him to make an inference about the mean of the parent population based on each of the two given sampling distributions to make sure that he understood the intended inference in this data-context. The following segments from the transcript show the sequence of the dialogue that led to the main question.
O: If you would like to make a prediction about the population based on each one of them

[Each one of the two sampling distributions given in Gym Task. Figure 6], what will be your prediction?

T: Do you want me to give like a number of what I would think?

O: Yeah, let me repeat the question. The question says, “What appears to be the population mean based on each plot?

T: So I guess for the first one [the spread out sampling distribution], the only thing I could do is find like the middle of all of those, which would be around like 56, just because there aren’t, there’s no cluster to look at really. There’s some [clusters] with two [repeated data values] but there’s some with two all the way at the beginning and some in the middle so I feel like that’s not really as reliable. Whereas the second one, I mean, I kind of have to do the same thing but I’d say probably more around like 52 or 54, just because that seems more balanced. So, I guess I’m kind of finding the means of each, like in my head, trying to find like the balance of each one. So, I mean they’re, I would say pretty much somewhere between 52 and 56 for both of them.

In the last response, Tanner jumped to the idea of comparing the appropriateness of the given sampling distributions for making inferences before I asked him to do so. He seemed to prefer the sampling distribution with a smaller range over the spread-out distribution just because it contained clusters that he could use to find the mean. Although he described the estimated mean of the spread-out sampling distribution as “that’s not really as reliable,” it is not clear if he was talking about the method of finding this mean or about this mean as an estimation of the
population mean because it came out of a spread-out sampling distribution. This issue led me to ask the next question.

O: Which one do you trust more?

T: I guess the second one, just because it was easier for me to find right away which, like what I think the mean would be of it. And like there are those clusters [Tanner pointed to the clusters at the middle of the sampling distribution with the smaller range as shown in Figure 6], but I guess making it easier doesn’t really make it more accurate. I don’t know. I’d still say the second one just because I see the clusters in it rather than just it being so spread out.

In contrast to his answer in the Bean Task, I argue that Tanner was not aware of the inverse relationship between sampling variability and the confidence in the inferences in this data-context. Although Tanner mentioned the word “cluster” in the above answer, he didn’t really use it to estimate the mean. (In fact, what he was actually doing was finding the median). By looking at the sampling distributions in Figure 6, one can notice that 52-54 lies in the central location of the data values. Moreover, there were no strong clusters in this interval.

In the following, I asked him which mean was easier for him to estimate. He gave a detailed answer in which he said explicitly that he preferred the spread-out sampling distribution over one with a smaller range. Strikingly this does not match his answer in the Gym Task.

O: But you didn’t use a cluster. You were talking about the mean.

T: Right.
O: And you have the same number of dots, by the way, in each of them [the two sampling distributions given in the Gym Task], you have 20 here, 20 there. So why do you think the second one is easier for you?

T: Because the data is more like, I don’t know. Maybe I would do the first one actually because like I was going, like I was saying before, the range makes me like verify, I think like there was like a good method of sampling done because there is such a range of means, rather than the second one is kind of just all pushed together. So, I guess the first one, even though it might not visually, like help me make the assumption of what the mean would be. But I feel like it is, like since it is such a variety, you can kind of find the mean of all of those and it will be your answer even though those like outliers will affect them. I feel like that’s kind of what they’re supposed to do so I mean, like if that many people in the sample said 34, then it should affect it like should affect your answer in finding the mean, I think.

The above answer illustrates that Tanner was thinking that spread out sampling outcomes are more representative of the parent population than clustered sampling outcomes because they give a variety of values and therefore a clearer picture about the shape of the population that is also made up of a variety of values. What Tanner missed here is that the goal of the sampling process in this task is to estimate a specific population parameter, the mean, rather than to estimate the shape of the whole population. In the following segment, I reminded him of a previous answer in which he preferred the spread-out sampling distribution.

O: So, you said you prefer the upper one [The upper sampling distribution in the gym context, Figure 6]. But you had a previous answer [In the Bean Task] in which you said
that I would like the outcomes to be similar. The outcomes here [...] are not similar!

You have kind of closer or similar [sampling outcomes].

T: Yeah, I know. I keep going back and forth because I like see good and bad things in both of them, but now I’m kind of leaning towards the first one [T refers to a set of data with more variability] being so spread out, I like more, because it seems like there was a better sampling method, just because of the variety of results and that’s what it would be like, I think, if you sampled the whole population, if you were able to like ask the whole population, I think it would be so spread out.

Reminding Tanner of his previous answer made him aware of the existence of some discrepancy. Yet, this awareness didn’t push him to rethink about his recent answer in which he prefers the spread-out sampling distribution. Instead, he stuck with this answer and justified it by employing the idea of the “sampling method” and its influence on the shape of the sampling distribution. I think he wasn’t so confident of his answer and therefore he tried to justify it by using any related notion that he could remember. It was easy for him to invoke the notion of the sampling method because he had class discussions about the different types of sampling methods during the two weeks that proceeded the interview. Statistically speaking, a better sampling method leads to less bias and therefore more clustered sampling distribution. Although the idea of sampling method is appropriate to mention and employ here, Tanner was not able to make use of it appropriately. What he missed here is that the data values in the intended sampling distribution, Figure 6, are the “means” of the selected samples rather than actual data values, and the intended estimation in this sampling situation is the mean of the population rather than the actual shape of the population. Although the data collected during the classroom activity was considered when I was conducting this analysis, it didn't include contextuality episodes that
could be highlighted in this analysis. Tanner stayed silent almost all the time during the activity and wrote brief responses on the worksheet. In these responses, he preferred large samples for making inferences because large samples better represent the population. Recall that the purpose of the classroom activity was to use TinkerPlots sampler to investigate the effect of increasing the size of the samples on sampling variability.

**Summary of Strands Two and Three: Selecting and Describing the Phenomena**

In the above, I have described two examples of contextuality related to two of the facets of the notion of sampling variability described in Chapter 5. These two facets are: (1) the effect of the size of the samples on the sampling variability, and (2) the inverse relationship between sampling variability and confidence in the inferences to be drawn based on the sampling distribution. The following chart, *Figure 12*, summarizes Tanner’s reasoning about these two facets of the concept of sampling variability across the Bean and the Gym Tasks.
In the next section, some elements in Figure 12 become refined, decomposed and reorganized. Systemizing the language employed in this next layer of analysis allows me to more precisely describe contextuality and hypothesize what is behind the phenomenon of contextuality of Tanner’s reasoning across data-contexts. This, however, is an interesting instructional implication per se. That is, systematizing my analytic language has made my conclusions less situated and more independent from the details of the present data-contexts, therefore, more generalized and applicable to a variety of learning contexts.
Re-describing the phenomenon of contextuality using Knowledge in Pieces (Analytical Strand Four)

As mentioned in Chapter 4, the Knowledge-in-Pieces (KiP) epistemology highlights diverse, fine-grained knowledge elements that can support inferring and reasoning about some notion often based on case study data of knowledge in-transition (Parnafes & diSessa, 2013), that is, describing at a moment-by-moment level what knowledge was being used and how it was changing. These characteristics of KiP make it a useful candidate as a theoretical framework for this dissertation study. Specifically, its fine-grained quality allowed a rigorous examination of processes of knowledge re-organization. Moreover, it enabled me to zoom in on the learning process and analyze the cognitive dynamics of the transitions that occurred in knowledge organization.

The metaphor of a complex system of knowledge elements of diverse kinds and functions in different data-contexts was employed in the analysis of the data in this phase. Moreover, the process of analyzing the data in this phase didn’t start with predetermined codes because of the unexpected knowledge elements in the PSTs’ systems of knowledge about the complex notion of sampling variability.

Similar to Wagner’s study, which also drew upon the KiP perspective, this section of the chapter will focus on Tanner’s understanding of some facets of the notion of sampling variability, offering an explanation of some of the contextualities and difficulties experienced by this PST as he reasoned with varying degrees of success about this notion across different data-contexts. What is different between this layer of the data analysis (analytical strand four) and the previous layer (analytical strands two and three), is the refined understanding of what knowledge resources Tanner was activating and offering conjectures for what is behind these differences (e.g., what was he attending to in each of the tasks and how did that impact what knowledge
resources were cued in that data-context). Refined understanding of contextuality can account for the occasions in which PSTs reason one way, and the occasions in which they reason another way.

**Disequilibrium about the effect of the size of the samples on the sampling variability**

In the Bean Task, Tanner noticed that one of the samples that he got using the small scoop had no red beans. Because it’s highly unlikely to get such a sample using the larger scoop, he inferred that large samples (scoops) will decrease the sampling variability. This noticing has led him to cue the knowledge element that “increasing the size of the samples decreases sampling variability.” Recall that *concept projection* stands for the reasoning strategies that enable the learner to recognize and apply the concept within some data-context. Therefore, cueing this knowledge element and applying it productively in this data-context indicates the construction of some concept projection associated with the facet of the sampling variability and associated with the effect of the sample size on the range of the sampling distribution. This doesn’t mean that this knowledge element is totally new for Tanner because it might be something he noticed in class when he used *TinkerPlots* to investigate the effect of the size of the samples on the sampling variability, but we can’t tell based on this short answer how robust was his understanding of this facet of sampling variability.

In view of KiP perspective, this concept projection is among the collection of concept projections that Tanner needs to construct (maybe in a wide span of data-contexts) and also *align* in order to construct expertise about this facet of sampling variability. Before we discuss the alignment across different data-contexts, let us see what this concept projection entails about the relationship between decreasing sampling variability and the range of the sampling distribution. To answer this question, I first asked what he meant by “sampling variability.” As shown in the
following excerpt, Tanner answered the question that I was planning to ask next which is about the effect of the sample size on the range specifically.

O: What do you mean by variability?

T: Like this \textit{the data values on the sampling distribution\textit{] ranges all the way from 0 to 36\%. I would expect it \textit{the data values\textit{ to be like not as much of a range in between there [Tanner pointed between 0\% and 36\%].

That he mentioned the “range” in his last answer provides evidence that his concept projection entails an awareness of the effect of increasing the size of the samples on the range in this data-context. \textit{Figure 13} summarizes the knowledge elements activated by the end of the Bean Task.

\textit{Figure 13}. Tanner’s knowledge system about the relationship between the size of the samples and sampling variability after solving the Bean Task
So far, Tanner has constructed a concept projection in the Bean Task that increasing the size of the samples will decrease the range of the sampling distribution. The question now becomes, can Tanner recognize and use this concept projection in a different data-context productively? The answer came shortly during the Gym Task when Tanner was asked, what might have caused the difference in spread between the two given sampling distributions? Based on his answer in the Bean Task, he was expected to mention the effect of the size of the samples as one of the potential reasons behind this difference between the two sampling distributions. Surprisingly he didn’t apply the previous concept projection in this data-context and argued that sampling method might have caused the difference between the two sampling distributions. We can’t consider this as evidence that his understanding of this facet of sampling variability is emerging because, statistically speaking, the sampling method is one of the factors that influence the shape of the sampling distribution. As mentioned previously in this chapter, Tanner had four sessions (100 minutes each) of class instruction across which sampling methods and their influence on the sampling outcomes were discussed. I conjecture that recognizing the relevance of the sampling methods in this task exhibits the existence of some concept projection associated with the sampling methods and their influence on the sampling outcomes. Within his knowledge system, it seems that the concept projection associated with sampling methods has a high cueing priority (e.g., it is the reasoning that is most readily available to Tanner). Probing his understanding of the relationship between the size of the samples and sampling variability requires a judicious strategy that draws his attention to this relationship without telling the answer. Therefore, I pushed him to reconsider his reasoning by drawing his attention to the fact that the size of the samples from the two cases in the Gym Task were not the same. This cued him to think about the size of the samples as an influential factor on the shape of the distribution.
Surprisingly he gave an answer that didn’t match his previous answer in the Bean Task. That is, he argued in the Gym Task that the sampling distribution with the wider range represents samples with larger sizes compared to the compact distribution with the smaller range. *Figure 14* illustrates the knowledge elements activated during the Gym Task.

![The Context of the Gym Task](image)

*Figure 14.* Tanner’s knowledge system about the relationship between the size of the samples and sampling variability after solving the Gym Task

This reasoning, however, is evidence that he couldn’t apply the previous concept projection productively in this data-context, therefore, his understanding of this facet of sampling variability is still rudimentary. In view of the KiP perspective, Tanner has constructed a new concept projection associated with the relationship between the size of the samples and the range of the sampling distribution in the Gym Task. This response indicates that his concept projection in the Bean Task was limited to the bean data-context and therefore it was hard for him to recognize its relevance to the gym data-context. Also, the context sensitivity of his first answer
might have impeded him from seeing similarities between the two questions. In other words, it was the sample with 0% red beans in the Bean Task that helped him infer the knowledge element and construct the concept projection in the data-context of the Bean Task. However, the Gym data-context had no similar sample with 0% (an extreme minimum) that might have helped him recognize the old concept projection and apply it in this data-context. A successful learning process is always reflected in the alignment of the concept projections associated with some notion. This apparently hasn’t been accomplished yet in the case of Tanner.

Thus far, Tanner has given two contrasting responses for two questions that entail the same statistical idea. These responses indicate different concept projections that are not aligned in his knowledge system. *Figure 15* illustrates the knowledge elements activated during each of the Bean and the Gym Tasks. So far, Tanner had not noticed any discrepancy between the concept projections that he has constructed in both data-contexts, therefore, *Figure 15* doesn’t show any bond or discrepancy between the two data-contexts.

*Figure 15*. Tanner’s knowledge system about the relationship between the size of the samples and sampling variability after solving each of the Bean and the Gym Tasks
To induce him to address this apparent contradiction, I reminded him of his answer in the Bean Task. He reaffirmed the original answer that the range will shrink but with a clear hesitation. He first said “it could grow” then changed his mind and reaffirmed the original answer and justified it using the same reasoning again, the 0% sample. He then tried to justify the increase of the range but ended his response by saying the range will shrink. This hesitation indicated some effort to align the two concept projections. Also, the reaffirmation of the previous answer might indicate that this concept projection still has high cueing priority even after constructing a new concept projection in the Gym Task. Tanner might have not seen the contradiction between these two concept projections clearly because he didn’t see a clear similarity between the two data-contexts.

O: How about the range?

T: I guess the range could grow. Well, actually, no, I think it will shrink because there’s less of a chance of you getting 0% of red beans with that cup than there is with that one. But then there’s also a bigger chance of you getting more than like four beans or 36% of red beans. So I could see it going either way. I guess it would shrink just because… but that’s just because I don’t think you’d have 0%, or you’d have less 0%.

In the above response, Tanner tried to apply the new concept projection that he constructed in the Gym Task by justifying the increase of the range of bean’s distribution. He thought that it’s more likely to get red beans in the larger scoop (sample) therefore the percentage of the red bean in the larger samples is more likely to be higher. Tanner was talking about the number of the red beans in each sample instead of the percentage of the red beans. Although this is a proportional reasoning idea, it’s of special importance to sampling variability because estimating population proportion is one of the common “point estimates” in statistics.
However, he then changed his mind and ended with the conclusion that the range will shrink. It wasn’t clear why he finally concluded that the range would shrink although he thought that the upper bound would increase. Perhaps he thought that the increase of the lower bound would exceed the increase of the upper bound and therefore the range would shrink.

When I asked why he thought that increasing the size of the samples would shrink the range in the Bean Task but expand it in the Gym Task he seemed not sure what to say. He just repeated his reasoning when responding to the Gym Task and emphasized that the range should grow in that case. It seems that my question was said with some surprise in a way that indicated to him that these two tasks are the same and therefore they should have the same answer. As a result, he experienced a disequilibrium.

O: Why you think that increasing the size of the samples here [the Gym Task] will not make it [the range] shrink?

T: Interesting. Because it’s more representative of the data. So, I guess it would grow because you’re asking more or you’re using more people in your sample, so you’re getting more variety in your answers. So, I guess it’s [selecting larger samples] more representative overall because you’re asking more people. So, I guess this would also, the range would also grow for this.

As shown in Figure 16, Tanner’s knowledge system still contains all of the knowledge elements and concept projections that he has constructed so far although he became aware that there should be something wrong in one of his answers because he got the impression from the tone of my question that they should be the same. So far, he doesn’t know even which of his answers should be changed because each of them is based on a concept projection that has a high cueing priority in his knowledge system.
Figure 16. Tanner’s knowledge system after cueing his attention to the equivalence of the Bean and the Gym Tasks

I tried to help him resolve this disequilibrium by pointing to the 0% in beans samples. I was hoping that this would help him see the concept projection that he constructed in the Bean Task as relevant to the Gym Task. Recall that the purpose of introducing new situations during the interview that might invoke learning is just a means of testing out my conjectures about the way a learner’s knowledge systems are organized; therefore, they are not intended to be learning tools during the interview. All of the bean samples were still on the table in front of Tanner. Because there was no 0% sample in the Gym Task, he didn’t think about the meaning of the 0% in the Gym Task at all but instead tried to resolve the disequilibrium by getting rid of the first answer. I argue that he hadn’t lowered the cueing priority of the concept projection that he constructed in the Bean Task because he didn’t judge or change the fundamental idea that he had built his concept projection on, which is the rare chance of getting 0% red beans using the large scoop.
In the interview segment below, he argued that decreasing sampling variability—as a result of increasing the size of the samples—wouldn’t necessarily mean decreasing the range but would rather mean getting more data values that would be closer to each other, not excluding the possibility of getting outliers that cause the range to increase. This answer, however, doesn’t represent a normative pattern of reasoning about this facet of sampling variability but it was convincing for Tanner.

T: Yeah, I think my original point [in the Bean Task] was that I would get more like similar, like instead of the range being from 0 to 36, I’d get more like similar, like a cluster, I guess you could say, of results with just a few outliers, rather than having this big of a range [Tanner pointed to the beans sampling distribution on the board, Figure 1]. So now I’m going back to my original answer to that. It would be smaller, I think.

As shown in Figure 17, Tanner has changed the concept projection that he has constructed while answering the Gym Task. Knowledge elements highlighted in bold type in Figure 17 were assumed to be actively used in the solution to the task. Tanner inferred a new knowledge element that the data values in the middle of the sampling distribution shrinks while the range grows as the size of the samples grows. He also deactivated the knowledge element that “sampling variability increases as the size of the samples increases” in favor of the knowledge element “sampling variability decreases as the size of the samples increases.” Tanner thought that he had removed the discrepancy by this answer, but I think one contrasting concept projection remained active in his knowledge system which is the idea that large scoops (samples) are not expected to scoop up outliers. Therefore, the lower bound of the range would increase as the sizes of the samples increased. Tanner had never negated this idea in any of his responses therefore I think this concept projection was still active in his knowledge system, but perhaps he couldn’t
recognize its relevance in the gym’s data-context. The above response underscores that he has tried to resolve the discrepancy. From the KiP perspective, resolving the disequilibrium means that he managed to align all of the concept projections associated with this facet of sampling variability across all of the available data-contexts.

![Diagram](image)

*Figure 17.* Tanner’s knowledge system after resolving the disequilibrium about the relationship between the size of the samples and sampling variability

An alignment should remove such contradictions and only keep the concept projections that are applicable across all of the available data-contexts. From the KiP perspective, this is part of a continuous learning process in which Tanner reduces the contextuality; he aligns his concept projections until he reaches a level at which he can align any concept projection smoothly without trouble. Although Tanner has experienced some disequilibrium in the above task, this doesn’t mean that a decline in his knowledge system has happened. More data-contexts that
entail this facet of sampling variability might help Tanner wring out this disequilibrium quickly because a new concept projection that supports one of the conflicting concept projections might be constructed.

**Disequilibrium about the effect of the sampling variability on the confidence in the inference**

Another sampling variability facet that Tanner wrestled to align his concept projections with was the inverse relationship between sampling variability and the confidence in the drawn inferences. In the following, I will use segments from the transcript to analyze this phenomenon from the KiP perspective. Early during the interview, Tanner was asked whether he would you like the sampling outcomes to be similar or different if he makes conclusions (inferences) about the percentage of the red beans in the parent population. His answer was:

T: I guess similar so that you can see, like, without these being similar, you wouldn’t have that cluster, so you wouldn’t be able to like see exactly where it’s going to be. But also without the differences, you wouldn’t be able to really tell how like extreme it is. Like in this one [Tanner pointed to the screen of the laptop which is shown Figure 11], there’s like 55% in between [Tanner pointed to the lower and upper bounds of the range then estimated the range which is 94% - 40% ≈ 55%] and I think that’s a pretty extreme like difference. So, I think mostly, like having the cluster in the middle is more beneficial than having the range, if that makes sense.

The first part of this response suggests that Tanner would like the sampling variability to be small in order to see clusters that help with making predictions. He continued by arguing that having a small range is also helpful for making a prediction. He clarified his response by pointing to the sampling distribution in Figure 11 and saying that the range 94% - 40% ≈ 55%, is
large and therefore not helpful for making predictions. Although he ended his response by saying “having the cluster in the middle is more beneficial,” he also has an appreciation for the range as an important factor. From a KiP point of view, this response indicated the activation of the two nested concept projections in his knowledge system which are, less sampling variability is desirable for making inferences about the parent population and within this concept projection, clusters are helpful because they help with determining the expected value. Figure 18 illustrates the concept projections constructed in Tanner’s knowledge system so far.

![Image](image.png)

*Figure 18. Tanner’s knowledge system associated with the inverse relationship between sampling variability and the confidence in the drawn inferences in the data-context of the Bean Task*

Statistically, this answer seems sufficient and appropriate for this question. However, from the KiP perspective, what is still needed is to affirm that Tanner can give equivalent answers for this question in other data-contexts. That is, can Tanner use this concept projection productively across different data-contexts?

In a later episode during the interview, Tanner was asked to answer a similar question in a different data-context. That is, he was asked which of the sampling distributions in the Gym Task he preferred if he made a conclusion about the parent population. His answer in the following segment underscores that he is hesitating to confirm that compact sampling distribution is more reliable.
T: I guess the second one [Tanner is referring to the lower sampling distribution in the Gym Task which is more compact], just because it was easier for me to find right away which, like what I think the mean would be of it. And like there are those clusters, but I guess making it easier doesn’t really make it more accurate. I don’t know. I’d still say the second one just because I see the clusters in it rather than just it being so spread out.

Neither of the sampling distributions in the Gym Task contained a clear cluster but one of them notably had a smaller range as shown in Figure 6. Although there were many small clusters around different values in the compact sampling distribution, Tanner wasn’t able to use the previous concept projection that compact sampling distribution is more reliable in this data-context without seeing a clear cluster. The hesitation in his answer may have indicated that he had experienced some disequilibrium caused by, on one hand, his recognition of the relevance of the previous concept projection and, on the other hand, his inability to apply this concept projection in the present context. This disequilibrium also indicated that his first concept projection about the inverse relationship between sampling variability and the confidence in the inference was narrow in a way that doesn’t consider the range but rather focuses on the cluster. It seems that he thought that a desirable sampling variability meant a clear cluster, like the beans sampling distribution, without any consideration for the range. For example, a sampling distribution with a large range and one cluster is more reliable to make inferences than any sampling distribution with no clusters (or with many clusters) but smaller range. Statistically, this is not quite accurate because the smaller the range, the more reliable is the sampling distribution for making inferences.

As he was trying to resolve the disequilibrium and use the previous concept projection in this new data-context, it seems that Tanner had noticed that the range of the compact distribution
is smaller than the other distribution. He tried in the above response to exploit this cued knowledge element to justify why the compact distribution was more reliable in the gym data-context by saying that it was easier to find the mean for this compact distribution.

The statement “making it easier doesn’t really make it more accurate” indicates that he has immediately deactivated the new knowledge element that he just cued in favor of his old activated knowledge element maybe because the old one has higher cueing priority in his knowledge system. With the deactivation of the new knowledge element, Tanner still had the same concept projections in his knowledge system compared to what he had constructed in the bean’s data-context. *Figure 19* illustrates Tanner’s knowledge system associated with the inverse relationship between sampling variability and the confidence in the drawn inferences after the first response in the gym’s context. The deactivated knowledge element is illustrated in gray.

![Figure 19](image.png)

*Figure 19.* Tanner’s knowledge system associated with the inverse relationship between sampling variability and the confidence in the drawn inferences after the first response in the gym’s data-context.

Tanner ended his answer by saying that he preferred the compact sampling distribution because it had more clusters, and this indicates that he managed to apply his previous concept projection in this new data-context regardless of the accuracy of his answer. Statistically
speaking, adding more data values (selecting more samples) to the sampling distribution will make a clear cluster around the expected value and this expected value equals the mean because the sampling distribution will become a normal (symmetric) distribution that has the mean right in the middle. To test the robustness of his concept projection that he managed to apply in two data-contexts so far, I reminded Tanner that each of sampling distributions in the Gym Task contains 20 data values then asked him why he thinks that the compact one is easier for making conclusions as long as they have the same number of data values. He answered as follows:

T: Because the data is more like, I don’t know. Maybe I would do the first one [the spread out sampling distribution in the Gym Task] actually because like I was going, like I was saying before, the range makes me like verify, I think like there was like a good method of sampling done because there is such a range of means, rather than the second one is kind of just all pushed together. So, I guess the first one, even though it might not visually, like help me make the assumption of what the mean would be. But I feel like it is, like since it such a variety, you can kind of find the mean of all of those and it will be your answer even though those like outliers will affect them. I feel like that’s kind of what they’re supposed to do so I mean, like if that many people in the sample said 34, then it should affect it like should affect your answer in finding the mean, I think.

In the above answer, Tanner has activated a new knowledge element that entails a discrepancy with the previous concept projection. He argued that the sampling distribution with a wider range is more reliable or representative of the parent population than the other sampling distribution because it indicates a better sampling method. He added that although a spread-out sampling distribution doesn’t visually help with estimating the expected value, it’s mean is a
good estimation for the expected value. By justifying the use of the mean as an alternative of the clusters for estimating the expected value, Tanner seemed to raise the contextual priority—the degree of confidence in the applicability of the concept projection in the relevant data-context—of the new concept projection. *Figure 20* illustrates the change to Tanner’s new knowledge system after answering the above follow-up question.

*Figure 20.* Tanner’s knowledge system associated with the inverse relationship between sampling variability and the confidence in the drawn inferences after the second response in the gym’s data-context

Before the follow-up question, Tanner seemed to be competent with this facet of sampling variability because he managed to apply it in two data-contexts with different degrees of productivity. His subsequent account, however, seemed to be deficient because it’s not statistically accurate. diSessa (1996) found that learners might provide both normative and non-normative reasoning when answering two similar questions in response to interviewer’s deliberate “shifts in attention” to different aspects of some concept. If Tanner has a stable concept projection, then he should have not changed it in response to my shifts in attention strategy.
Chapter Summary

Data analysis presented in Chapter 6 aimed, in particular, at answering the second research question in this study “How are PSTs’ reasoning processes about sampling variability sensitive to data-context?” The analysis provided an answer to this question by demonstrating that contextuality was richly displayed in Tanner’s reasoning about sampling variability and illustrating the difficulties and the contextualities experienced by Tanner. Although the follow-up and discussion questions used during the interview entailed similar facets of sampling variability across tasks, Tanner clearly did not see them that way and gave different answers to similar questions about two facets of sampling variability across two of the data-contexts— the Bean and the Gym Tasks. He also justified his answers differently across these data-contexts. These two facets of sampling variability were: (1) the effect of the size of the samples on the sampling variability, and (2) whether or not more or less sampling variability is desirable. The novelty of the analysis presented in this chapter is manifested in its fine-grained description of the difficulties and the high degree of contextuality in a participating PST’s reasoning about sampling variability. The data analysis in this chapter has revealed that sampling variability is a complex and multi-faceted notion that requires the learner to construct and align many concept projections across many data-contexts.
CHAPTER 7: DISCUSSIONS AND IMPLICATIONS

In this chapter, I looked across the analytic chapters and connected them to the research questions. Also, I situated the findings of these two chapters in the broader issues in the field of statistics teacher education. With the recent focus on the concept of sampling variability in influential policy reports such as the Common Core State Standards (CCSS-M, 2010) and Guidelines for Assessment and Instruction in Statistics Education Report, (GAISE, 2007), elementary/middle grades PSTs should acquire the specialized content knowledge needed to teach this concept successfully. This specialized content knowledge, however, is different than the specialized content knowledge that high school teachers need in order to teach sampling variability. That is, although both middle and high school curricula include sampling variability, high school students are expected to use advanced statistical tools to describe sampling variability such as the standard deviation that are beyond the reach of young learners. Moreover, high school students are expected to connect sampling variability with confidence intervals and testing hypotheses.

Having established that there are two levels of specialized content knowledge associated with sampling variability, a natural next question is “What should elementary/middle grades PSTs know about sampling variability?” Sampling variability in elementary/middle grades needs to be focused on the foundational ideas that underpin sampling variability. For example, the overall purpose of selecting samples is to make inferences about the population. Based on the results of my two Pilot Studies, I found that the PSTs I interviewed in my dissertation study seemed to have an intuitive sense that large samples were always better. It might, therefore, seem surprising that PSTs have any difficulty understanding a fundamental idea such as the behavior of the sampling variability when the size or the number of the samples increases. However,
Kahneman & Tversky, (1972) concluded that “the notion that sampling variability decreases in proportion to sample size [the law of the large numbers] is apparently not part of man's [sic] repertoire of intuitions” (p. 444). The law of large numbers is key to learning the concept of sampling variability, however, this concept is broad and developing a sufficiently complete understanding of it requires an awareness of many related statistical ideas. To take one example, the importance of which also came to light in pilot work: studying sampling variability, for example, requires selecting or simulating the selection of many small samples and developing a sampling distribution. Therefore, learners should be aware of the validity of selecting many small samples and its equivalence to selecting one large sample. Learning this idea becomes more important to beginning learners who tend to intuitively prefer large samples. According to CCSSM (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010), the first-time students use repeated small samples is when they study sampling variability in grade seven. Therefore, elementary/middle PSTs need to be aware of all the aspects associated with selecting many small samples in order to teach it successfully when they start their teaching career. In the following, I revisit the research questions addressed in this study, discuss my findings, and elaborate on possible contributions to the field of statistical education.

The First Research Question

The range of statistical ideas associated with the concept of sampling variability has not yet been fully explicated in the literature, especially with respect to elementary and middle grades PSTs who do not have access to many advanced statistical ideas. Therefore, the most important ideas for PSTs to learn about sampling variability would be those ideas that are challenging for them but also within the scope of the elementary and middle grades school
One of the primary goals of my dissertation study was to identify ways that PSTs reason about sampling variability and in particular to understand potential sources of difficulty. Thus, the first research question addressed in this study was “What non-normative ideas do PSTs invoke in reasoning about facets of sampling variability?” Chapter 5 aimed at answering this question by identifying any non-normative patterns of reasoning that the PSTs participating in my study exhibited as they worked on the tasks of the interview. The analysis in Chapter 5 provides a partial replication of prior research findings (e.g. Leavy 2010; De Vetten, Mooney, et al, 2014) showing that PSTs face some challenges when reasoning about sampling variability. In the current study, each of the seven PSTs used non-normative reasoning to justify their answer at least once during the interview. Also, there was a notable similarity across their reasoning processes that allowed me to identify patterns of non-normative reasoning about specific facets (aspects) of the concept of sampling variability. These facets of sampling variability were: (1) Whether or not selecting many small samples then combining them into one large sample is equivalent to finding the sample statistic for each sample separately, (2) Whether or not selecting many small samples then combining them is equivalent to selecting another large sample with the same number of subjects, (3) The effect of the sample size on the range and clusters in the data and (4) Whether or not more or less sampling variability is desirable.

The non-normative patterns of reasoning identified in Chapter 5 are similar to some of the ways of thinking about sampling variability identified in Pfannkuch’s (2008) framework. The analysis in this dissertation extended Pfannkuch’s work by clarifying and elaborating some of these ways of thinking—referred to as “images” in Pfannkuch’s study.

The “image of concept of random process” identified in Pfannkuch’s (2008) work involves an awareness of the underlying random process of selecting samples and how random
samples behave in terms of a particular data-context. It also involves an awareness of the sampling method and its effect on sampling outcomes. However, as Pfannkuch describes it, this image doesn’t consider the fundamental idea of the equivalence between selecting small samples and noticing a pattern, on one hand, versus combining these small samples into one large sample and then calculating the sample statistic for this large sample, on the other. Also, Pfannkuch’s “image of concept of random process” doesn’t consider another fundamental idea: the equivalence of selecting multiple small samples and then combining them, on one hand, versus selecting one large sample, on the other hand. These two fundamental ideas identified in the data analysis of the PSTs’ reasoning processes (cf. Chapter 5) underpin and rationalize selecting multiple small samples. This is important for understanding the concept of sampling variability because it leads to building a sampling distribution and understanding the logic of the law of the large numbers. With that being said, the equivalence of these two sampling strategies was not obvious to the participating elementary and middle grades PSTs. The non-normative pattern of reasoning associated with the equivalent sampling strategies suggests it should be considered in Pfannkuch’s (2008) framework for the ways of thinking about sampling variability under the “image of concept of random process.”

Because the primary goal of inferential statistics is to generalize from a sample to a population, the inference becomes more accurate if the sample size is large. For instance, when the sample size is small, it is possible to obtain a sample that contains many large values (making the sample statistic large) or many small values (making the sample statistic small). For larger samples, however, it is more likely that the values are equally spread between large and small, and any extreme values are balanced out with the other values to produce a more central value for the sample statistic. Therefore, selecting larger samples makes sampling outcomes closer to
each other. If these sampling outcomes were presented in a sampling distribution, then the range of this sampling distribution would be smaller. One of the ways of thinking concerning sampling variability identified in Pfannkuch’s (2008) study was the “image of intuitive confidence interval.” Pfannkuch describes this image in terms of creating images of sample statistics intervals for different sample sizes. Although this facet highlights the effect of the sample size on the confidence intervals, there was no clear mention in the description of this facet for the exact effect of the sample size on the range of the sampling distribution. It might be because Pfannkuch’s framework was initially intended to be used with high school students who are expected to use confidence intervals in their reasoning. The analysis from Chapter 5 revealed clear patterns of non-normative reasoning in PSTs’ responses about the effect of the sample size on the range of the sampling distribution. For example, some of them argued that increasing the size of the samples will increase the range of the sampling distribution. These patterns highlight the need for explicitly addressing the effect of the sample size on the range of the sampling distribution in Pfannkuch’s framework if we use it to accurately describe elementary and middle grades PSTs' ways of thinking about sampling variability.

Statistically, less sampling variability reflects a better sampling method, making the sampling distribution more reliable for making inferences about the population. Although this idea seems intuitive to PSTs, the data analysis in Chapter 5 revealed that some PSTs have an emerging understanding of the overall purpose of the sampling distribution and the relationship between its shape and the trustworthiness of the drawn inferences. That is, PSTs don’t see the inverse relationship between the range of the sampling distribution, on one hand, and the trustworthiness of the drawn inference or conclusion, on the other. This study found that some PSTs preferred sampling distributions with wider ranges to make inferences because they
thought that these sampling distributions better represented the diversity that naturally exists in any population. This conclusion is in line with the findings of some previous studies (e.g., Rubin, Bruce, & Tenney, 1991; Gil & Ben-Zvi, 2011) which showed that learners struggle with balancing between “representativeness” and “sampling variability.” For instance, Rubin et al., (1991) found that overreliance on sampling representativeness leads the learners to think that the sample tells us everything about the population. At the opposite extreme, overreliance on sampling variability may also lead learners to think that the sample tells us nothing useful about the population.

Based on the results of this study, I would argue that PSTs should build an understanding of the balance between “representativeness” and “sampling variability” by, (1) drawing their attention to noticing the difference between sampling distributions built from samples with different sizes and (2) maintaining the connection between sampling variability and inference at all times. In light of this, it is notable that Pfannkuch’s (2008) framework did not consider students’ thinking about the relationship between the shape of the sampling distribution and the trustworthiness of the inference. Given that the present study is focused on elementary and middle grades PSTs, I advance the proposal that PSTs’ ways of thinking concerning this relationship be addressed as a new category in Pfannkuch’s (2008) framework. I also suggest strengthening the connection between sampling variability and making inferences by redesigning Pfannkuch’s (2008) framework to take a pyramid shape where “making an inference” is positioned at the top of the pyramid as shown in Figure 21. My intention is to call attention to the overall purpose of selecting samples (making inferences about the parent population). In fact, the act of making inferences builds upon all of the facets of sampling variability, and if these facets are not well understood, PSTs may never fully develop their abilities to make sound inferences.
The above discussion of PSTs' non-normative reasoning about the notion of sampling variability highlights the need for explicitly clarifying learners’ reasoning about these four identified facets in Pfannkuch’s (2008) framework especially if we use this framework with elementary/middle grades PSTs. That is, these PSTs are expected to teach an informal version of sampling variability that focuses on the foundational ideas that underpin this concept. Figure 21 illustrates the suggested adaption of some of the images identified in Pfannkuch’s (2008) framework. The images that I am elaborating are “Intuitive confidence interval” and “Concept of random process.” In addition to elaborating and refining these two images that already exist in Pfannkuch’s framework, I suggest adding the image “Relationship of sampling distribution and inference” to the framework.
Figure 21. Framework for ways of thinking about sampling variability
More than mathematics, learning statistics is inextricably tied with data-contexts because the data are not just numbers but numbers with a context, therefore, data-context should be considered in any statistical reasoning (Cobb & Moore, 1997). Thus, the role of the data-context was vital to the data analysis in Chapter 5. Discussing sampling variability with the participating PSTs in three different data-contexts during the interview raised to the surface some contextualities experienced by some of these PSTs as they reasoned with varying degrees of success about the same facet of sampling variability across different data-contexts.

Based on the findings of Wagner (2006) and the principles of the coordination class model (diSessa & Sherin, 1998; diSessa & Wagner, 2005; diSessa, Sherin, & Levin, 2016), and in light of the findings of the data analysis in Chapter 5, I hypothesized that PSTs' reasoning about sampling variability, in particular, might be highly contextual. Thus, an in-depth analysis of clear case in which a PST exhibits multiple instances of contextuality in their reasoning processes about sampling variability was of particular interest for this study.

**The Second Research Question**

Chapter 6 aimed, in particular, at answering the second research question in this study “How are PSTs’ reasoning processes about sampling variability sensitive to data-context?” The analysis developed in Chapter 6 answered this question by demonstrating that contextuality was richly displayed in Tanner’s reasoning about sampling variability and illustrating the difficulties and the contextualities experienced by Tanner. Statistically, the follow-up and discussion questions used during the interview entailed similar facets of sampling variability across tasks. However, Tanner clearly did not see them that way and gave different answers to similar questions about two facets of sampling variability across two of the data-contexts—the Bean and the Gym Tasks. These two facets were: (1) the effect of the size of the samples on the sampling
variability, and (2) whether or not more or less sampling variability is desirable. The novelty of the analysis in this chapter centered on the fine-grained description of the high degree of contextuality in a participating PSTs’ reasoning about sampling variability.

Concerning Tanner’s contextualities and difficulties about the effect of the size of the samples on the sampling variability, the analysis in Chapter 6 revealed that Tanner had constructed a concept projection in the Bean Task that increasing the size of the samples will decrease the range of the sampling distribution. However, he couldn’t recognize and use this concept projection productively later in the data-context of the Gym Task. Moreover, he constructed a new concept projection in the Gym Task that contradicted this concept projection constructed in his reasoning about the Bean Task. This indicated that his concept projection in the Bean Task was limited to the Bean data-context and therefore it was likely hard for him to recognize its relevance to the Gym data-context. Moreover, my data analysis revealed that the context sensitivity of his first concept projection might have impeded him from seeing similarities between the two questions.

Tanner also exhibited clear contextualities and difficulties about the relationship between the shape of the sampling distribution and the trustworthiness of the inferences to be drawn based on the sampling distribution. Data analysis in Chapter 6 also revealed that he has constructed two contrasting concept projection across the Bean Task and the Gym Task during the interview. That is, he argued in the Bean Task that less sampling variability is desirable for making inferences about the population, while he argued in the Gym Task that spread out sampling distribution is more reliable for making inferences.

Later when he became aware of the discrepancy between his answers in both of the above cases, he experienced a disequilibrium because he didn’t know which of his answers should be
changed. This indicated that each of the two contrasting answers in both of the cases was based on a concept projection that has a high cueing priority in his knowledge system, therefore, he couldn’t align them. In both cases, Tanner tried to resolve the discrepancy by raising the contextual priority—the degree of confidence in the applicability of one of these concept projections in the other context. From the KiP perspective, resolving the disequilibrium means that the learner managed to align all of the concept projections associated with this facet of sampling variability across all of the available data-contexts. An alignment should remove such contradictions and only keep the concept projections that are applicable across all of the available data-contexts. From a KiP perspective, this is part of a continuous learning process in which Tanner reduces the contextuality until he reaches a level at which he can align any concept projection smoothly without trouble. Tanner experienced some disequilibrium in his knowledge system twice and faced alignment difficulties as he was working on the tasks of the interview. What he needs is an exceedingly wider span of different data-contexts using sampling variability before he develops expertise (coordination class) for this concept because thinking in different contexts involves constructing new concept projections with a continuous alignment process (Wagner, 2006).

The combination of productivity of Tanner’s knowledge elements and the contextuality of these knowledge elements used in data analysis of Chapter 6 illustrates a key feature of analyses informed by Knowledge in Pieces. As the episodes presented in this chapter show, some knowledge elements or concept projections—even if they are not usually cued—can be helpful when brought to Tanner’s attention in one particular data-context because they might be combined and coordinated; therefore, they might likely to be used productively by Tanner in an increasingly wider span of data-contexts.
As mentioned before, data analysis on the PSTs’ reasoning in Chapter 5 has provided a partial replication of prior research findings showing PSTs face some challenges when reasoning about sampling variability. The data analysis reported in Chapter 6 results in a much stronger caution: even detailed reasoning about sampling variability in one data-context may not be sufficient to say that a PST has a complete understanding of this notion. Moreover, successful reasoning about one facet of sampling variability doesn’t guarantee that the learner can provide successful reasoning about the other facets of this complex notion. This finding is consistent with the prior work on reasoning in mathematics, physics, and statistics done from the Knowledge in Pieces perspective (e.g., diSessa, 1993, 2015; Levin, 2018; Izsák & Jacobson, 2017; Wagner, 2006; 2010) in which contextuality of reasoning processes is one of the central assumptions. However, the unique aspect contributed by this study is the empirical grounding and detailed specification of contextuality in PSTs’ reasoning about sampling variability, with little formal background in statistics.

This study was further unique in the sense that it employed two research strands to answer the posed research questions. That is, the use of Pfannkuch’s (2008) framework concerning the ways of thinking about sampling variability form the field of statistics education along with Knowledge in Pieces from the learning sciences has opened new research avenues concerning sampling variability that take into consideration detailed aspects of this concept. One goal of the data analysis in Chapter 6 was to show that coordination classes model (diSessa & Sherin, 1998; diSessa & Wagner, 2005; diSessa, Sherin, & Levin, 2016) can be used effectively to investigate the multi-faceted concept of sampling variability. By doing so, more research from the learning sciences arena might investigate learning this concept in depth.
Implications for Instruction

The results of this study lead to implications for statistics teacher education and possibly for precollege teaching. To begin with, some of the participating PSTs argued that combining many small samples together into one large sample then finding the sample statistic for this large sample is more reliable for making inferences than finding the sample statistic for each of them separately then looking for any pattern in the resulting data values. To help PSTs appreciate the latter sampling strategy, which is vital for understanding sampling variability, I suggest teacher preparation curricula should use multiple tasks that include each of these two sampling strategies and for PSTs to compare the outcomes and unpack what is going on in each strategy. For example, the Voting Task and the Bean Task could be activities done in class with physical and computer-based sampling simulations. Dynamic data visualization software such as TinkerPlots might be helpful as they can simulate the selection of the samples with different sizes and present the sampling outcomes. Similarly, other PSTs preferred selecting one large sample over selecting many small samples then combining them together to form one large sample even if the size in both cases was the same. Again, teacher preparation curricula should explicitly include authentic tasks in which PSTs experience collecting multiple samples and try each of these two sampling strategies and compare the outcomes. These issues might also appear in the school setting and these authentic tasks might be used effectively because they don’t involve any advanced statistical tools that are beyond the reach of middle grades’ students.

Based on the findings of the data analysis in Chapter 5, PSTs need the opportunity to think about the effect of the size of the selected samples on the range of the sampling distribution. Furthermore, they need to be encouraged to think about the effect of the size of the samples on both the upper and the lower bounds of the range especially if they work on physical
sampling simulation tasks. More importantly, studying sampling variability in K-8 preservice teachers content courses should be tied to informal inferential reasoning. That is, PSTs need to keep in mind as they reason about sampling variability that the overall goal of studying sampling variability is to help with making accurate inferences about the parent population. Therefore, they need to be aware of the relationship between the shape of the sampling distribution and the trustworthiness of the inferences to be drawn. I also suggest that special attention ought to be given in curricula design to building understanding and a careful balance between sampling representativeness and sampling variability because it seemed that PSTs who preferred spread out sampling distributions to make inferences have over-relied on sampling representativeness in their overall thinking about the situation. Based on the findings in Chapter Six, sampling variability is a complex and multi-faceted concept where true understanding requires PSTs to encounter multiple data-contexts that entail the concept. Moreover, these data-contexts need to include all of the sampling variability facets discussed in Chapter Five with follow up and reflection questions that encourage the PSTs to think about the similarities between the different data-contexts.

As mentioned in the theoretical framework chapter, contextuality of reasoning processes has been documented across a large number of studies in many disciplines. The question becomes what can we do to meaningfully support PSTs' reasoning through instruction. For this, it is essential to know about the ways in which student reasoning is contextually grounded. We need to talk to our PSTs, gather data, and look at the way they reason. From that, we build appropriate instructional materials. When PSTs experience disequilibrium in their knowledge system because of the way that they construe a data context and the features they attend to, the disequilibrium doesn't always result in a more normative organization of their knowledge
systems. Careful follow-up questions need to be asked by the teacher to know what exactly happened in their knowledge systems as a result of the experienced disequilibrium. The case of Tanner clearly exhibited that although he wrestled with disequilibrium for some time, this didn't help him get determine which was the statistically accurate reasoning pattern. While experiencing disequilibrium may be an important instructional moment, it is far from clear yet how to most effectively leverage such an experience or whether this is always the most desirable intervention.

**Limitations and Future Work**

Based on the findings of the current study and the limitations, the following recommendations are made. This dissertation study is only a start to the process of understanding PSTs' reasoning about the concept of sampling variability as a concept that underpins their IIR and the contextuality phenomenon that was reported in this study. Undoubtedly, an extended study with more participants utilizing more data-contexts would be valuable in that it would have the potential to provide a more detailed, complete, and generalizable picture of the nature of PSTs’ knowledge and reasoning process about sampling variability as well as a deeper understanding of the contextuality phenomenon that was reported in this study. A future study could allow for comparisons among PSTs’ responses from this current study to PSTs’ responses from different classrooms and altered sampling variability tasks that entail new physical or computer-based sampling simulation data-contexts.

This current study was largely exploratory regarding the way PSTs’ intuitive (pre-instructional) knowledge systems concerning sampling variability are organized in addition to testing out the conjectures about the contextualities they exhibited as they reason about this concept across different data-contexts at a moment in time (relatively early in the semester).
Therefore, the tasks used during the interview were not intended primarily to be learning tools. Because these tasks were new to the PSTs, I think they might invoke learning. Future research could benefit from the use of these tasks to investigate the development of PSTs’ understanding of sampling variability. Future research might track the development in their understanding at different points in time during the semester to see what patterns in their reasoning might look like after having more experience and instruction around sampling variability. This could involve collecting more data from the assessment tools used in the probability and statistics content course and by interviewing the PSTs again at the end of the semester. Future work that follows this dissertation study should focus more on the ways by which we can support PSTs in constructing tasks that involve different data contexts (including the three tasks used in this study) as similar.

All of the participating PSTs were successful in describing the effect of increasing the number of samples on the range. That is, they all argued that the range might grow as the number of the samples increases which is an accurate answer statistically. Those PSTs, however, might give non-normative reasoning about the effect of increasing the number of samples on the range in different data-contexts. Therefore, it might be the next research step. Any future research that catches PSTs' non-normative reasoning about the effect of the number of the samples on the range would easily extend this facet. The current study was focused on the specialized content knowledge that PSTs need in order to teach sampling variability successfully. I suggest that further research should be carried out to address the pedagogical content knowledge that PSTs also need in light of the difficulty and contextuality involved in learning this concept.

Each of the three tasks used during the interview involved opportunities to discuss different facets of sampling variability and connect it to the overarching idea of IIR. However,
there were some minor differences between these tasks due to the used sampling methods. For example, sampling during the Bean Task was without replacement whereas sampling during the Voting Task was with replacement. None of the participating PSTs attended to this difference. Future work might take this difference into consideration by returning the bean' samples to the container every time to make the tasks more consistent. Future research might also check the equivalence between all of the follow-up and reflection across all of the three tasks.

Because of the small number of participating PSTs in this study, generalizations can’t reliably be made from this data set about the distribution of the identified patterns of non-normative reasoning across all elementary and middle grades PSTs. Further research with large number of participating PSTs would needed in order to investigate the follow-up question of how widespread these non-normative patterns of reasoning are among PSTs. However, this wasn't among the purposes of this study which by part aimed at describing the nature of these patterns of non-normative reasoning in PSTs’ intuitive (pre-instruction) knowledge systems. This study accomplished this goal by clearly highlighting and describing four different non-normative patterns of reasoning about facets of sampling variability. This study also accomplished its second goal which is describing in fine-grained a clear case of contextuality. Increasing the validity of the data analysis presented in this study was by presenting ample examples from the transcriptions of the interviews to support each step or conclusion in my data analysis. That is, there was no hidden data that was used to make any of the conclusions in this study. However, all the data analysis was conducted by one person which is the only reason that I am aware of that might have lowered the validity in this study. This limitation will be taken into consideration when I publish this study by asking a second coder to reconduct this data analysis from a fresh perspective to ensure a higher level of validity. Future research might compare the findings of
the data analysis presented in Chapters Five and Six with another analysis for the same data using the Structure of Observed Learning Outcomes (SOLO) taxonomy (Shaughnessy, 2007). The SOLO taxonomy is a general hierarchical model of cognitive development in which a learner progresses through five modes of thinking: sensory (motor), ikonic (images), concrete symbolic, formal, and post formal. This taxonomy has seen a great deal of attention recently, particularly in statistics education research. The hierarchical organization of SOLO involves a somewhat different hypothesis about knowledge organization and learning processes (as progressing through stages in order) to the systems perspective of Knowledge in Pieces that does not a priori assume such a pattern of development, allowing instead for non-linear learning pathways to be documented.

**Closing Remarks**

I began the Pilot Studies for this line of research with the intention of investigating PSTs' informal inferential reasoning. Through a further review of the literature related to the role of sampling variability in developing IIR (e.g., Pfannkuch, Arnold, & Wild, 2015; Garfield & Ben-Zvi, 2007) and the role of the context in statistics reasoning and thinking (e.g., Cobb, 2007; Wgner, 2006; Wild & Pfannkuch, 1999), I started to get a clearer picture of what work lay in front of me. I became more focused on studying sampling variability as one of the most important concepts that underpin IIR. At the end of the two Pilot Studies, I had developed two research questions along with a research plan to answer each of them. Through the course of analyzing the dissertation data, these two research questions have undergone slight changes. While I was unable to find any examples of other people attempting to develop a framework that helps with describing K-8 preservice teachers reasoning about sampling variability, few examples exist in the literature of people proposing frameworks for high school students (e.g.,
Pfannkuch, 2008; Shaughnessy, 2007). The statistical background of the high school students in Pfannkuch’s study was comparable to the statistical background of the elementary/middle school PSTs in my study, and so it was reasonable to expect that the framework would provide a solid foundation for my research.

In spite of that, there was a need for testing out the appropriateness of Pfannkuch’s (2008) framework for PSTs. In Chapter 5 of this dissertation study, I presented my analysis of the participating PSTs' reasoning about sampling variability and how I used this analysis to elaborate Pfannkuch's (2008) framework. The elaborated version of this framework served well the fine-grained analysis of the role of the context in PSTs' reasoning about sampling variability in Chapter 6. That is, the contextuality analysis took into consideration facets of sampling variability that haven't been considered in Pfannkuch's (2008) framework among them is the connection between sampling variability and the trustworthiness of the inferences to be drawn about the parent population. The contextuality analysis presented in Chapter 6 is a significant step toward more research that investigates PSTs' reasoning about the multi-faceted concept of sampling variability. Showing a clear case of contextuality using the principles of the coordination classes model will open doors for similar studies that might build upon the methods and findings of this study.
REFERENCES


Appendix A: HSIRB Approval Letter

Date: January 9, 2018
To: Mariama Levin, Principal Investigator
    Omar Abu-Ghalyoun, Student Investigator for dissertation
From: Amy Naugle, Ph.D., Chair
Re: HSIRB Project Number 17-12-26

This letter will serve as confirmation that your research project titled “Developing K-8 Preservice Teachers’ Informal Inferential Reasoning Using a Sampling Simulation” has been approved under the expedited category of review by the Human Subjects Institutional Review Board. The conditions and duration of this approval are specified in the Policies of Western Michigan University. You may now begin to implement the research as described in the application.

Please note: This research may only be conducted exactly in the form it was approved. You must seek specific board approval for any changes in this project (e.g., you must request a post approval change to enroll subjects beyond the number stated in your application under “Number of subjects you want to complete the study”). Failure to obtain approval for changes will result in a protocol deviation. In addition, if there are any unanticipated adverse reactions or unanticipated events associated with the conduct of this research, you should immediately suspend the project and contact the Chair of the HSIRB for consultation.

Reapproval of the project is required if it extends beyond the termination date stated below.

The Board wishes you success in the pursuit of your research goals.

Approval Termination: January 8, 2019
Appendix B: Participant Consent Form

Western Michigan University
Department of Mathematics

Principal Investigator: Dr. Mariana Levin, Department of Mathematics
Student Investigator: Omar Abu Ghalyoun
Title of Study: Developing K-8 Preservice Teachers’ Informal Inferential Reasoning Using a Sampling Simulation

You are invited to participate in a research study titled “Developing K-8 Preservice Teachers’ Informal Inferential Reasoning Using a Sampling Simulation” in connection with your MATH 2650 course. This consent document will explain the purpose of this research project and will go over all of the time commitments, the procedures used in the study, and the risks and benefits of participating in this research project. Please read this consent form carefully and completely and please ask any questions if you need more clarification.

What are we trying to find out in this study?
In Math 2650, Probability and Statistics for the Elementary/Middle School Teachers, students explore the mathematical content of probability and statistics, use technological tools to support and extend their thinking, and work in collaborative groups to peer- and self-assess. The primary purpose of this study is to investigate how elementary/middle school mathematics PSTs develop their statistical reasoning when they explore the sampling variability using the dynamic software TinkerPlots and how does this affect the meaning of the experimental probability they have.

Who can participate in this study?
All students enrolled in the Math 2650 classes that Dr. Christine Browning will teach during the 2018 academic year are encouraged to participate in the study, “Developing K-8 Preservice Teachers’ Informal Inferential Reasoning Using a Sampling Simulation”. The only inclusionary criterion is the student need to be enrolled in a M2650 class that Dr. Christine Browning will teach during the 2018 academic year. Students who self-select to not participate in this study will be excluded without any negative impact on them. Among all of the students who choose to participate in this study, only four students will be selected randomly by the researcher to be the focus of this study. There are no consequences for not participating in the study. The identity of those who choose to participate or not will not be made known to course instructors until after the course grades are submitted for the current term when formal data analysis will take place.

Where will this study take place?
Part of this study will take place in the classroom of Math 2650 on the main campus of Western Michigan University. You may also choose to participate in an interview component of the study. If you choose to do the interviews, they will take place on campus at a mutually convenient place, and would occur during the semester. To participate in the study, you DO NOT have to choose to be interviewed. That is a separate component of the study.
What is the time commitment for participating in this study?
The time commitment for participating in this study is the length of the current semester since the data collected consists of the tasks you complete during the term. If you choose to participate in the interview component of the study, those will also take place during the term. The interview itself will last no longer than 30 minutes.

What will you be asked to do if you choose to participate in this study?
All data collected will come from course tasks that are part of the normal classroom practice, therefore you are not asked to do anything beyond what is expected of any student enrolled in a M2650, other than granting permission for the researcher to copy your work. Your name will be removed prior to any copying of your work to ensure anonymity in the data analysis. Code names will be created and used to permit examining your (anonymous) statistical thinking across the semester.

If you choose to participate in the interview component of the study, you will be asked to describe your thinking in a completed task. The interview will last no longer than 45 minutes, use your code name to ensure confidentiality, take place on campus at mutually convenient place between you and the interviewer, and would occur during the semester. The interviewer will be a member of the research team and not the course instructor. To participate in the study, you DO NOT have to choose to be interviewed. That is a separate component of the study.

What information is being measured during the study?
This study is examining your understanding for an important skill in statistics which is the informal inferential reasoning when you explore the sampling variability using the dynamic software TinkerPlots and the impact of developing this skill on your understanding of the basic concepts of probability. To this end, your work from course tasks and your responses during the interviews and will be analyzed to see how your statistical thinking changes and to assess the impact of the tasks and technological tool on your thinking.

What are the risks of participating in this study and how will these risks be minimized?
There are no additional risks to participating in this study, other than those that may occur as part of your daily routine as an undergraduate student. If you choose to participate in the interview component of the study, you may experience some minor discomfort that can occur during interviews (e.g. mild stress owing to the one-to-one setting). When the interviewer observes such discomfort, s/he will provide wait time for re-composure and remind you that you are able to halt the interview at any point in time without consequence. You may choose to remove yourself from the study or the interview at any time with no course grade penalty.
What are the benefits of participating in this study?
What is learned from this study may benefit the future course instructors of M2650 at Western Michigan University, future Math 2650 students and, through published results, inform researchers and course instructors at other institutions of ways to provide the best possible preparation for beginning elementary/middle school teachers in the area of probability and statistics while utilizing digital tools.

Are there any costs associated with participating in this study?
There are no costs associated with participating in this study other than the time commitment given if you choose to participate in the interview component. Those who participate in the study are completing the work expected of any student enrolled in a Math 2650 class that Dr. Christine Browning will teach during the 2018 academic year.

Is there any compensation for participating in this study?
Those participating in the interviews will get $15 gift card and may gain more insight into their own thinking and learning processes. Also, they might benefit from the tasks which are very similar to their classroom tasks and homework.

Who will have access to the information collected during this study?
Only Dr. Mariana Levin and Omar Abu-Ghalyoun, will have access to the information collected during this study. All data will have names removed prior to copying and code names assigned. Video data collected (the interviews) will use code names when necessary to insure confidentiality. If any collected information is used during a public presentation, your identity will be kept confidential through the use of these code names or other pseudonyms. Dr. Christine Browning won’t know who has grant permission for us to copy his/her course assignments nor who has participated in the interviews until after final grades have been turned in for the course.

What if you want to stop participating in this study?
You can choose to stop participating in the study at any time for any reason. You will not suffer any prejudice or penalty by your decision to stop your participation. You will experience NO consequences either academically or personally if you choose to withdraw from this study. To stop participating email Dr. Mariana Levin at mariana.levin@wmich.edu or Mr. Omar Abu-Ghalyoun at omar.ghalyoun@wmich.edu and indicate that you no longer want to participate in the study.

Should you have any questions prior to or during the study, you can contact the primary investigator, Mariana Levin at mariana.levin@wmich.edu or Omar Abu-Ghalyoun at 269-501-4720 or omar.ghalyoun@wmich.edu.
You may also contact the Chair, Human Subjects Institutional Review Board at 269-387-8293 or the Vice President for Research at 269-387-8298 if questions arise during the course of the study.

This consent document has been approved for use for one year by the Human Subjects Institutional Review Board (HSIRB) as indicated by the stamped date and signature of the board chair in the upper right corner. Do not participate in this study if the stamped date is older than one year.

Please continue to provide an indication of your participation levels.

__________________________________________________________

IF YOU AGREE TO PARTICIPATE, PLEASE CHECK ONLY ONE OPTION

___I have read this informed consent document. The risks and benefits have been explained to me. I agree to take part in this study, but I do not want to be interviewed.

___I have read this informed consent document. The risks and benefits have been explained to me. I agree to take part in this study and I am willing to be interviewed.

__________________________________________________________

Please Print Your Name

__________________________________________________________

Participant’s signature Date
Appendix C: Beginning of Class Interview Protocol—Fall 2017

Interview Protocol

Research title: Developing K-8 preservice teachers’ informal statistical inference in a dynamic software environment

Interviewer: Omar Abu-Ghalyoun (Ph.D. student at WMU)

Interviewee:

Time and place:

Thank you for agreeing to be interviewed. This interview is designed to examine your statistics and probability thinking. We can use your responses to better understand how students can best learn the statistical content knowledge needed to be an effective teacher. Your answers will remain confidential and will not be used for grading purposes by the instructor. If during the interview, you decide to stop the questioning and not continue, there is no penalty. Thank you for your willingness to participate in the interview.

Your background:

Tell me about yourself,

1. What’s is your age?

2. What’s your major, what year are you?

3. Have you taken a statistics class before Math 2650? If so, what was it?

Research Questions:

Question 1:

a. If you roll this die, what will be the possible outcomes?

b. In probability, people frequently say this die is ‘fair.’ What does this word mean for you in the probability context?
c.  Jon rolled a similar dice 10 times and got 4,2,4,6,2,1,5,2,6. He claimed that this die is loaded because #3 has never appeared. Do you agree with Jon?

d.  Do you have a way by which we can test whether a die is loaded or not?

e.  If you roll a die once, what is the probability (chance) of getting a 2? a 5? Explain.

**Question 2:** Imagine the test scores for a group of 1000 middle students in the large school district. The test scores for a random sample of ten students from this class are shown in the dot plot [below]. (*This question has been adapted from Zieffler et al., 2007*)

(Zieffler et al., 2007)

Now, consider a random sample of 25 students drawn from this school district. Try to imagine what THAT graph might look like. Explain your reasoning.
Question 3: The mean average speed in miles per hour and length of flight in miles were recorded for 27 airline flights. The scatterplot of these data is shown below. (This question has been adapted from https://locus.statisticseducation.org)

Make an inference (conclusion) about the relationship between flight length and average speed.
Appendix D: Beginning of Class Questionnaire— Spring 2018

Background Survey and Initial Thoughts About Samples and Sampling
* Required

Background Information
This first section asks some basic background questions related to your prior experience with statistics.

1. Name *

2. Email *

3. Age *

4. Have you taken a statistics class before Math 2650? If so, what was it? *

Initial Thinking About Samples
In the next sections, you will be asked to think about three scenarios. Throughout Mathematics 2650, you will gain tools for thinking about these kinds of situations. The purpose here is to get your initial thoughts about these situations. Please take some time to consider each situation and share any thinking you have that supports your conclusions about the situations.

Rolling Die

5. Jon rolled a die (six faces) 10 times and got: 4, 2, 4, 6, 2, 2, 1, 5, 2, 6. He claimed that this die is loaded because #3 has never appeared. Do you agree with Jon? Try to include all the relevant reasons for your answer. Do you have a way by which we can test whether a die is loaded or not? *
6. On a particular day, 65% of the births in hospital A were female, and 30% of the births in hospital B were female. On the next day, which hospital is more likely to have more female births than male? Try to include all the relevant reasons for your answer.

Test Scores

7. Imagine the test scores for a group of 1000 middle students in large school district. The test scores for a random sample of 7 students from this district are 92, 84, 80, 77, 95, 87, 90. Now, consider another random sample of 7 students drawn from this school district. What might be their scores? Predict the 7 values and explain your reasoning.
Appendix E: Interview and In-Class Tasks

TASK 1

Name: ................................. Date: .................................

Voters Task (growing sample size)

Proposition 223 (Prop 223) requires each of California's school districts to limit certain administrative costs to 5 percent of all federal, state and local funds received, beginning in 1999-2000. Presumably, this means that 95 percent of all funds will go to the actual education of children in K-12, including the salaries of classroom teachers. It also requires each school district, beginning in 1998-99 to tie its annual budget to specific outcomes, generally related to improvements in student performance.

Source:
https://www.calvoter.org/voter/elections/archive/98primary/caljournal/measures/prop223.html

Prop 223 TinkerPlots sampler (find it in the TinkerPlots folder) will be used in this task to simulate randomly selecting samples from California state voters who have been asked whether they would vote for or against Prop 223. Before you start, it’s important for you to understand what the number 10 on the icon exactly means, and what does the icon do. Ask your shoulder neighbor or raise your hand. In this task, you will work in groups and each member of the group needs to do one of the following:

I. Change the sample size to 100 then run the sampler 30 times. Copy your results to the following table.

II. Change the sample size to 500 then run the sampler 30 times. Copy your results to the following table.

III. Change the sample size to 1000 then run the sampler 30 times. Copy your results to the following table.

IV. Change the sample size to 2000 then run the sampler 30 times. Copy your results to the following table.

V. Change the sample size to 3000 then run the sampler 30 times. Copy your results to the following table.
<table>
<thead>
<tr>
<th>Percents of the people who would vote for Prop. 223 in a sample</th>
<th>Number of the samples having the same percent as the first column.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Person 1: 30 samples each of size 100</td>
</tr>
<tr>
<td>48%–50%</td>
<td></td>
</tr>
<tr>
<td>51%–53%</td>
<td></td>
</tr>
<tr>
<td>54%–56%</td>
<td></td>
</tr>
<tr>
<td>57%–59%</td>
<td></td>
</tr>
<tr>
<td>60%–62%</td>
<td></td>
</tr>
<tr>
<td>63%–65%</td>
<td></td>
</tr>
<tr>
<td>66%–68%</td>
<td></td>
</tr>
<tr>
<td>69%–71%</td>
<td></td>
</tr>
<tr>
<td>72%–74%</td>
<td></td>
</tr>
<tr>
<td>75%–77%</td>
<td></td>
</tr>
</tbody>
</table>

1. What is the sum of the numbers in each column? Why would you expect to get this sum?

2. What inference would you draw from all the samples of size 1000 regarding Proposition 223?

3. What inference would you draw from all the samples of size 3000 regarding Proposition 223?

4. What does a sample of 3000 give you that a sample of 1000 does not?
5. What general pattern do you see on the table? What does your pattern tell you about the effect of sample size?

**TASK 2**

Name: ……………………………... Date: ………………………...

Voting Task (growing number of samples)

Similar to the previous task, Prop 223 TinkerPlots sampler (find it in the TinkerPlots folder) will be used in this task to simulate randomly selecting samples from California state voters who have been asked whether they would vote for or against Prop 223.

1. Run the sampler once to create one sample of 15 responses. What percentage did you find for the two outcomes (For and Against)? What do these percentages mean?

2. In this activity, we will collect many samples of size 15. Do you expect to see an approximately similar percentage of the people who were for Prop 223 across these samples?

3. Based upon this sample of 15, what inference (conclusion/prediction) could you make about the percent of the people who would vote for Prop 223 in California state?

4. How confident are you in the accuracy of your inference?

5. Create another sample of 15 responses. Does the new sample make you any more (or less) confident about your inference? Explain your thinking?

6. I would like you to collect 50 samples each of size 15 and see if the percentage of the people who were “For” Prop 223 is consistent across all of the samples. What do you notice?

7. Can anything make the outcomes of the samples closer to each other?
8. Now let us test the effect of collecting 50 more samples. I will take a picture of this history of results plot and keep for a comparison when you test out this idea. I would like you to generate another 50 samples (the total now is 100 samples). What do you notice?

9. How did the distribution of the data change? Compare to the photo I took of the previous distribution.

10. Has the variability of the percentages reduced or increased as we collect more samples?

11. Give an interval of values in which you are fairly certain that the actual percentage of the voters who would vote “For” the proposition lies.

12. Before you come to the interview, I collected 50 samples twice. My results are shown in the following two sampling distributions.

13. What appears to be the percentage of the people who would vote for Prop. 223 based on each sampling distribution?

14. Which of sampling distribution you feel more confident to use to make a conclusion about the population?

15. Given that both sampling distribution has the same number of samples, what might have caused these differences between them?
16. Instead of collecting 50 samples each of size 15, assume we selected 750 people once at random. Do you think we are getting more confident in the percentage of the people who would vote "for" the proposition in the true population?

17. Please change the size of the sample to be 750 then collect one sample. What do you think now?

TASK 3

The Bean Task

This bowel contains 800 white beans and 200 red beans. Assume that these numbers are unknown, and you would like to estimate the percentage of the red beans only in this container using some sampling technique.

1. Fill these two scoops with beans. What percentage of beans do you expect to get in each scoop?

2. Assume that we have only this small scope. Can we use it to estimate the percentage of the red beans? How?

3. Let us collect the same number of beans using these two scoops and put them in two piles outside the box. Which pile do you trust more as a sample that represents the population?
Appendix F: Interview Protocol Used in Fall 2018

Interview Protocol

Thank you for agreeing to be interviewed. This interview is designed to examine your statistics and probability thinking. We can use your responses to better understand how students can best learn the statistical content knowledge needed to be an effective teacher. Your answers will remain confidential and will not be used for grading purposes by the instructor. If during the interview, you decide to stop the questioning and not continue, there is no penalty. Thank you for your willingness to participate in the interview.

Interview Task 1: The Bean Task

This container contains 800 white beans and 200 red beans.

1. If we fill these two scoops with beans, what percentage of red beans do you expect to get in each scoop? Why?

2. Which scoop would you like to use to estimate the percentage of the red beans in this box? Why?

3. Assume we have only this small scoop. Can we use it to estimate the percentage of the red beans? How?
   - If a PST suggests using the small scoop repeatedly to select a large sample outside the box, I will ask,
     - Assume you can’t keep more than one sample outside the container (return it before you collect a new sample). Is it still beneficial to use this small scoop?
   - If a PST suggests using the small scoop to collect multiple samples and record the outcomes then I will ask,
     - How many samples will you collect?
     - Will the percentages of the red beans be similar across these samples?
       - If the answer is no then I will ask,
         - Then why do you want to repeat collecting non-similar samples?
         - How can you trust many inconsistent outcomes?
       - If the answer was yes, then I will ask,
         - What do you mean by ------ (similar, close, the same)?
• Why you expect the percentages to be similar across these samples.
• Could you please collect few samples using this small scoop to convince me more than the outcomes will be similar?

4. What would you expect to see if we used this large scoop to select these 10 samples instead of the small one?

5. Assume that I changed the role and allowed you to collect the same number of beans using these two scoops and put them in two piles outside the box. Which pile do you trust more as a sample that represents the population?
   • I will say thought-provoking statements such as:
     o Remember, this pile is made up of small samples!
     o Remember, both piles contain the same number of beans!

Interview Task 2: Voting Task (growing number of samples)

Proposition 223 (Prop 223) requires each of California's school districts to limit certain administrative costs to 5 percent of all federal, state and local funds received, beginning in 1999-2000. Presumably, this means that 95 percent of all funds will go to the actual education of children in K-12, including the salaries of classroom teachers. It also requires each school district to tie its annual budget to specific outcomes, generally related to improvements in student performance.

In this task, a TinkerPlots sampler will be used to simulate randomly selecting samples from California state voters who have been asked whether they would vote for or against Prop 223.

1. I will run the sampler once to create one sample of 15 responses. What percentage did we get for the two outcomes (For and Against)? What do these percentages mean?
   • Here I will explain the purpose of the history display.

2. What conclusion can you draw from this sample about the percentage of the people who voted for Prop 223 in California state?
   • How confident are you in your conclusion?
   • What do you mean by ------ (any certainty or uncertainty term)?
   • What makes you confident? not confident?
   • Are there any other reasons that make you confident or not confident?
3. I will select another sample of 15 people. What percentage do you expect to get?
   - Why do you expect this percentage?
   - Why it’s hard to expect any percentage?
   - Did the results from the sample we drew before influence the answer you gave this time? Why?

4. Does the new sample make you any more (or less) confident about your inference?
   - What is it that makes you less or more confident?
   - If the PST argued that the new sample will make her more confident because the percentages of the two samples are similar, I will ask:
     - Do you think the percentage in the next sample will also be similar? Why?
   - If the results are very different in the two samples, I will ask “What do you make of the difference between the two samples?”

5. I will select 20 samples each of size 15. What do you expect to get?
   - If the answer was “different percentages,” I will ask:
     - Why?
     - Do you think you can say anything about the interval that the percentages will fall between? Which values?
   - If the answer was “similar percentages”, I will ask:
     - Why?
     - Does the similarity mean getting a specific percentage?
     - What percentage is this?
   - If the answer didn’t include anything about the percentage, I will ask:
     - Can you say anything about the percentages of the people who were for Prop 223 across these samples?”

6. (After selecting the samples, I will ask) what do you notice now?
   - What do you mean by -------- (vary, spread OR spaced out)?
     - What might have caused this variation?
   - What do you mean by -------- (clustered, centered around some value OR close to each other)?
   - What might have caused this cluster?
   - In general, do you think it is beneficial to keep a record of the history of sampling outcomes when we make conclusions about the population? why?

7. Would there be anything you could do to change the shape of the history display so that you feel more confident about your prediction?
   - Is it the only thing?
8. What do you think will happen to shape of the history display as we select more samples?

9. Suppose I select 100 samples in place of 20. What do you think the final graph will look like?

I will take a picture of this history of results plot and keep for comparison later.

10. Now let us select 80 more samples. The total now is 100 samples. What do you notice?

   • Compare with the photo I took of the previous history of results plot.
   • Has the shape of the plot changed?
   • What aspects of the plot have changed?
   • Why do you think the shape plot has changed? hasn’t changed?
     o If the PST uses the term “sampling variability” in any of the previous questions, then I will use it in the rest of the conversation instead of describing it informally.
   • Is the change you have noticed beneficial for making conclusions about the population? Why?

11. If you make a conclusion about the population based on some samples, would you like the sampling outcomes to be similar or different (sampling variability to be high or low)? Explain your answer.

12. Could you give an interval of values in which you are fairly certain that the actual percentage of the people who voted “For” the proposition lies?

   • How did you determine this interval?
   • If some PST suggests the use of some measures of center, I will ask: why do you think that ____ (measure of center) is appropriate to use here?

13. Instead of selecting 100 samples each of size 15, assume we selected 1500 people (the same total) just once. Should we be more confident about our conclusion than in the case of selecting 100 samples of size 15?

I’m going to give you now another task to think about.

**Interview Task 3: The Gym Task**

The question “What is the typical time spent at the gym?” is being investigated by selecting many samples from a population of 800 gym members. Displayed below are two different dot plots of sample means calculated from random samples of the population.
1. Describe the differences between the two dot plots.

2. What appears to be the population mean based on each plot?

3. Which dot plot are you more confident using to answer the statistical question? Explain your answer.

4. Given that both plots have the same number of samples, 20 samples, what might have caused the differences between the two plots?
   
   - What if I told you the sample sizes are different? Does that make a difference?