Informal statistical inference is the process of making inferences based on data without using formal statistical procedures. The purpose of this study is to examine how teachers use technology to solve an informal inference problem, and to determine their levels of reasoning about the relationship between expectation, variation, and sample size.

Fifty-seven pre-service and in-service teachers across three institutions in the United States wrote documents describing their problem-solving strategies for a specific task. These documents ranged in length from one page (including screenshots) to fifteen pages. This study utilized a case study approach that considered three different technologies, Probability Explorer, Microsoft Excel, and Fathom, to be three cases.

A framework to determine teachers’ levels of reasoning about the relationship between expectation, variation, and sample size was modified from the framework described by Watson, Callingham, and Kelly (2007). The modified framework outlined hierarchical levels of reasoning that ranged from Inconsistent to Explanatory Comparative. Teachers were coded at specific levels based on their responses and how they used simulated data to support their conclusions.

The findings from this study indicate that teachers who use Probability Explorer tend to simulate fewer repeated samples than teachers who use Microsoft Excel or Fathom, primarily because of constraints within the technology. More repeated samples were associated with higher levels of reasoning. The types of conclusions teachers formed also
appeared to be related to their choice of technology because each technology choice offered slightly different graphical representations, as well as different statistical functions. The majority of teachers were coded at the highest reasoning levels, but a substantial number of teachers still operated at the lower levels of reasoning.
Teachers’ Use of Technology in Solving an Informal Inference Problem

by

Jenna Marie Rice

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Master of Science

Mathematics Education

Raleigh, North Carolina

2013

APPROVED BY:

__________________________________
Dr. Herle McGowan

__________________________________
Dr. Allison W. McCulloch

__________________________________
Dr. Hollylynne S. Lee
Chair of Advisory Committee
Jenna Rice was born and raised in New Bern, North Carolina. She is the daughter of Frank and Kathleen Rice, and the younger sister of Jeff. Jenna graduated from New Bern High School in 2007 with honors, and went on to attend North Carolina State University for her undergraduate career. She graduated summa cum laude from N. C. State in 2011 with a B.S. in Mathematics Education and a B.S. in Statistics. After participating in research opportunities as an undergraduate in both statistics and mathematics education, Jenna decided to pursue an M.S. in Mathematics Education and began her graduate career during the summer of 2011.

Jenna currently teaches at Lee Early College High School in Sanford, NC. She has taught high school math there for one and a half years. Jenna has taught Geometry, Algebra 2, Advanced Functions and Modeling, and she is excited to teach Statistics this year. She enjoys teaching at LEC because of her wonderful colleagues and her clever and funny students.

In the future, Jenna hopes to pursue a Ph. D. in Mathematics Education with a focus on statistical learning at all grade levels. Until then, Jenna is committed to becoming a better teacher each year, and she hopes to make a positive impact in her community by serving in public education.
I would like to thank all of the teachers I have had the privilege of learning from throughout my life. From Pre-K to graduate school, my teachers have been invaluable in my growth as a life-long learner. Thank you for instilling confidence in my mathematical abilities, and for pushing me to be a better writer. Thank you for showing me the joy of learning, and that perseverance is the key to success. Thank you for the years of service you have committed to your community. Thank you. I hope to pay it forward.

I would like to express my gratitude to my advisor, Dr. Hollylynne Stohl Lee for her support and guidance throughout this process. The ability to bounce ideas off of someone who is so knowledgeable and to ask for advice or suggestions has been priceless. Also, I would like to thank Dr. McCulloch and Dr. McGowan for their suggestions and patience during this project, and for their willingness to serve as committee members.

I would also like to thank my family for believing in me, and for supporting me in my endeavors. Thank you for caring about my success in school without demanding unachievable perfection. I feel lucky to have a family so supportive of my ambitions, even when they seem to come out of nowhere. To my brother, Jeff, thank you for motivating me to reach the high bar you set before me.

For my friends who are teachers, thank you for working hard and making a commitment to education. The resources and inspiration you provide help make me a better teacher. For my friends who are not teachers, thank you for listening when I incessantly talk about the struggles and joys of teaching. Your support and kindness matter so much.
The research reported in this thesis was partially supported by the National Science Foundation under Grant Nos. DUE 0442319 and DUE 0817253 awarded to North Carolina State University. Any opinions, findings, and conclusions or recommendations expressed herein are those of the author and do not necessarily reflect the views of the National Science Foundation. The researcher also thanks the mathematics teacher education faculty and graduate students that were part of this NSF-funded project who assisted in collection and organization of data across several institutions. Thank you, too, to the many pre-service mathematics teachers that contributed their work that comprised the data for this study.
# TABLE OF CONTENTS

LIST OF TABLES ........................................................................................................... ix

LIST OF FIGURES ......................................................................................................... x

CHAPTER 1: INTRODUCTION ......................................................................................... 1

  Statement of the Problem .......................................................................................... 5

CHAPTER 2: LITERATURE REVIEW .............................................................................. 6

  Informal Statistical Inference .................................................................................. 6

  Frameworks for Informal Statistical Inference ......................................................... 7

  Samples .................................................................................................................... 8

  Sample Size .............................................................................................................. 9

  Sampling Variation .................................................................................................. 12

  Sampling Variation with Respect to Expectation ....................................................... 14

  Sampling Distributions ............................................................................................ 16

  Sample Collection ................................................................................................... 19

  Simulations ............................................................................................................... 20

    *Probability Explorer* .......................................................................................... 22

    *Fathom* ............................................................................................................... 25

    *Microsoft Excel* .................................................................................................. 28

  Graphing Calculator ................................................................................................. 30

  The Teacher’s Role .................................................................................................. 31

  Research Questions ................................................................................................ 33
LIST OF TABLES

Table 1      A Table Describing Watson et al.’s (2007) Coding Scheme for Variation and Expectation ..........................................................15
Table 2      Coding Scheme for Initial Expectations for the Proportion .....................40
Table 3      Coding Scheme for Variation with Respect to Sample Size .....................42
Table 4      Types of Conclusions Formed with Descriptions .........................................47
Table 5      Top Justifications for Choosing Probability Explorer ............................56
Table 6      The Reasoning Levels for Probability Explorer Users in Part (a) ...............58
Table 7      The Reasoning Levels for Probability Explorer Users in Part (b) ............59
Table 8      Frequency Table for Types of Samplers vs. Types of Conclusions for Probability Explorer ..........................................................63
Table 9      Top Justifications for Choosing Excel ......................................................65
Table 10     The Reasoning Levels for Excel Users in Part (a) .....................................66
Table 11     The Reasoning Levels for Excel Users in Part (b) .....................................68
Table 12     Frequency Table for Types of Samplers vs. Types of Conclusions for Excel ..........................................................71
Table 13     Top Justifications for Choosing Fathom .....................................................73
Table 14     The Reasoning Levels for Fathom Users in Part (a) ..................................74
Table 15     The Reasoning Levels for Fathom Users in Part (b) ..................................75
Table 16     A Frequency Table for Type of Sampler by Levels for Comparing Variations ........................................................................83
Table 17     A Frequency Table for Type of Sampler by Type of Conclusion ...............86
Table 18     Comparing Levels of Reasoning in Part (a) and Part (b) .........................88
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Designing an Experiment in <em>Probability Explorer</em></td>
<td>23</td>
</tr>
<tr>
<td>2</td>
<td>The Home Screen in <em>Probability Explorer</em> Showing the Results of Simulations</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>Simulating Random Integers in <em>Fathom</em></td>
<td>26</td>
</tr>
<tr>
<td>4</td>
<td>Collecting Measures and Generating Dot Plots in <em>Fathom</em></td>
<td>27</td>
</tr>
<tr>
<td>5</td>
<td>Generating Random Numbers in <em>Excel</em></td>
<td>28</td>
</tr>
<tr>
<td>6</td>
<td>“Count If” Command in <em>Excel</em></td>
<td>29</td>
</tr>
<tr>
<td>7</td>
<td>Generating Random Integers in a Graphing Calculator</td>
<td>30</td>
</tr>
<tr>
<td>8</td>
<td>The Informal Inference Task Teachers Solved</td>
<td>37</td>
</tr>
<tr>
<td>9</td>
<td>An Organization of the Problem-Solving Pattern and Related Questions</td>
<td>39</td>
</tr>
<tr>
<td>10</td>
<td>Two Data Tables of the Simulated Results of Two Samples in <em>Probability Explorer</em></td>
<td>49</td>
</tr>
<tr>
<td>11</td>
<td>Two Tables of the Simulated Results of Two Samples (Size 387) in <em>Excel</em></td>
<td>50</td>
</tr>
<tr>
<td>12</td>
<td>A Dot Plot in <em>Fathom</em> of the Number of Hats for 100 Simulated Samples (Size 93)</td>
<td>51</td>
</tr>
<tr>
<td>13</td>
<td>A Spreadsheet in <em>Excel</em> Showcasing Formal Procedures</td>
<td>52</td>
</tr>
<tr>
<td>14</td>
<td>Number of Repeated Samples Teachers Took Using <em>Probability Explorer</em></td>
<td>61</td>
</tr>
<tr>
<td>15</td>
<td>Number of Repeated Samples Teachers Took Using <em>Excel</em></td>
<td>69</td>
</tr>
<tr>
<td>16</td>
<td>Dot Plot of Number of Hats Given Away on Friday (93) in <em>Fathom</em></td>
<td>77</td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

In the current age of access to volumes of data, it is increasingly important that we are able to analyze and make inferences appropriately. Every day we read about studies citing statistics, or making claims about how one program or prescription is better than another. Policies that affect our entire society are often based on the results of experiments, observational studies, and surveys. Because of this, it is vital that all citizens have a basic level of statistical literacy that allows them to ask insightful questions about how data was collected and what techniques were used for analysis instead of blindly accepting another person’s conclusions as fact. By teaching students how to interpret and evaluate various types of data, teachers play an essential role in educating students to make them better citizens in society (Garfield & Ben-Zvi, 2008).

With the introduction of the Common Core Standards (CCSSM, National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010), mathematics educators will be expected to teach big statistical ideas beginning in sixth grade and continuing throughout high school. For example, some of the topics students will be expected to learn are: sampling variability, describing distributions, using random sampling to make inferences, comparing two populations, and making inferences based on experiments. Before the Common Core Standards were introduced, the National Council of Teachers of Mathematics (NCTM, 2000, 1989) focused an entire strand on data analysis and probability. One part of this strand was to “develop and evaluate inferences and predictions
that are based on data” (NCTM, 2000, p. 248). Specifically, high school students were expected to “use simulations to explore the variability of sample statistics from a known population and to construct sampling distributions,” and “understand how sample statistics reflect the values of population parameters and use sampling distributions as the basis for informal inference” (NCTM, 2000, p. 324). Based on these popular standards, statistical inference is a major topic that requires our attention.

Statistical inference has generally been regarded a difficult topic for students to understand – resulting in the need for learning about informal inferential reasoning first (Zieffler, Garfield, delMas & Reading, 2008; Makar & Rubin, 2009). Zieffler et al. (2008) and Makar & Rubin (2009) have suggested frameworks to help support future research on informal inferential reasoning, and offered examples of activities that help students learn about informal inference. These frameworks are designed to be broad in nature, allowing for activities at the elementary school level, or as complex as high school tasks. According to Makar and Rubin (2009), informal inferential reasoning in statistics is “the process of making probabilistic generalizations from (evidence with) data that extend beyond the data collected” (p. 83). Students can engage in an informal inference task by designing an experiment, collecting data, and making inferences using their data as supportive evidence.

One model for statistical investigations is the PCAI Model (Graham, 1987; Friel, O’Conner, & Mamer, 2006). The PCAI Model is a process for statistical investigations in which students pose questions, collect data (usually from randomized experiments), analyze the data, and make interpretations. This type of model would work well for informal
inference tasks because students are involved in collecting and analyzing the data, and then they must either interpret salient features of the data or draw inferences. The teacher’s role is to design or choose open-ended tasks that will be appropriate for his/her students. Additionally, the teacher should ask focused questions to keep the students engaged in task and attentive to the data being collected. While this sounds simple enough, teachers often struggle to ask questions that will deepen students’ understanding of the statistical topic they are learning (Lee & Mojica, 2008; Leavy, 2010).

In addition to choosing appropriate tasks and asking deeper questions, teachers also need to have the necessary content knowledge of the statistical topics embedded in informal inference. First, teachers need to know about samples, and that randomly selected samples accurately reflect characteristics of the population (Konold & Higgins, 2003). Second, teachers need to be aware of the relationship between expectation and variation, and how that relationship is affected by sample size (Watson, Callingham, & Kelly, 2007). Finally, teachers need to know that different samples yield different statistics, and that appropriate variation is difficult to intuitively predict (Watson & Kelly, 2004).

A major facet of informal statistical inference is to make inferences based on data. It is important for students to know how the data must be collected in order to reach unbiased conclusions. According to Rossman and Chance (1999), “statistical inference applies only to situations where sample data have been selected from a population or process or where experimental subjects have been randomly divided into treatment groups” (p. 302). In order to build student’s informal inferential reasoning, tasks should be constructed that let students
design experiments and collect data through simulation. After gaining experience with hands-on simulations, such as flipping a coin or rolling a dice, it is more efficient to use computer simulations (Drier, 2001; Rossman & Chance, 1999; Ben-Zvi, 2000). The use of simulations is also suggested by NCTM (2000) and CCSSM (2010).

In addition to having the necessary statistical knowledge, teachers also need to know how to use the technologies, particularly for conducting simulations, which are available for their classrooms. One of the main facets of building informal inferential reasoning is the requirement that students make arguments based on data. When all students are presented with the same data, they may not realize that results vary from sample to sample. Thus, it is important that students have the opportunity to design experiments and explore the results of simulations; this can help students understand that, although they may not have the same results as another student, the way in which they formed their conclusions may be similar. In order for these opportunities to take place, teachers need to be able to teach their students how to use the available technologies, and explain what the different features represent (Drier, 2001). For example, if a teacher shows his/her class how to randomly generate numbers on a calculator, they would need to discuss what each number would represent in the context of their data. By having an understanding of the technologies available in the classroom, the teacher can help ensure the success of his/her students as they collect data.

Although informal inference is not an easy topic to teach, it is crucial that teachers endeavor to teach it to the best of their abilities. Teachers need to be comfortable with both the statistical topics and the available technology in order to guide their students through an
informal inference task. By collecting simulated data for a real-world context, students can begin to observe the differences that occur between samples, and they will be able to determine which statistics are typical, and which would be considered unusual. Teachers can help guide this process by modeling problems, and by direct questioning or small group discussions. This will help lay the conceptual foundations students will need when they eventually encounter formal statistical inference.

**Statement of the Problem**

The purpose of this study is to examine how teachers engage in an informal inference task, and the ways in which their technology choice aids or constrains their endeavors. More research needs to be done on informal inference tasks in general, and while mathematics teachers do not represent the entire population of students, their strategies for solving problems could potentially highlight unforeseen misconceptions. Additionally, by considering the ways in which the teachers use different technologies, we can make suggestions to develop more intuitive and powerful technologies for future use, as well as improve professional development programs for learning statistical topics with technology. This study seeks to explore the mathematics teachers’ understanding of informal inference topics, develop ideas for future research, and offer recommendations for teacher education programs in informal inference topics.
CHAPTER 2
LITERATURE REVIEW

Informal statistical inference is a relatively new research topic that encompasses many statistical topics and has varying levels of complexity. The purpose of this review of the literature is to consider the research that has already been done and determine a cohesive way of viewing the major facets of informal inference. This literature review distinguishes formal and informal statistical inference, identifies common statistical ideas within informal inference problems, describes the role of technology in solving these problems, and considers the part that teachers play in teaching these topics.

Informal Statistical Inference

Formal statistical inference is notoriously difficult to teach which is one reason why informal statistical inference is now being taught at the 6-12 grade levels. According to Wild, Pfannkuch, Regan, and Horton (2011), the term ‘statistical inference’ refers “to the territory that is addressed by confidence intervals, critical values, p-values, and posterior distributions.” (p. 250). In contrast, informal inferential reasoning is defined by Zieffler, Garfield, delMas, and Reading (2008), “as the way in which students use their informal statistical knowledge to make arguments to support inferences about unknown populations based on observed samples.” (p. 44). While formal statistical inference topics are typically first taught in an AP Statistics class or an introductory-level college statistics class, informal inferential reasoning can and should be taught beginning in elementary school through high school. The tasks used can vary in complexity according to the students’ prior statistical
knowledge. Throughout the investigation of these tasks, students can build an understanding of samples, statistical measurements, variability, distributions, etc. Formal statistical inference requires not only an understanding of these big ideas, but also how to use what is observed to make conclusions. In order to develop an intuitive understanding of these processes, informal statistical inference plays a key role in building the conceptual groundwork that students will need.

*Frameworks for Informal Statistical Inference*

Several theoretical frameworks exist that help define and outline informal statistical inference. The Informal Inferential Reasoning (IIR) Framework, created by Zieffler et al. (2008), highlights three main ideas: making judgments or predictions based on data without using formal methods, using prior statistical knowledge, and making final claims about the population using data from samples (p. 45). Makar and Rubin (2009) also state the importance of making generalizations about the population with supporting evidence from the data, but they add the importance of using probabilistic language to indicate a level of certainty about conclusions. Vallecillos and Moreno (2002) determined four major constructs needed to engage in statistical inference: the relationship between populations and samples, the inferential process, sample sizes, and sampling types and biases. While these topics were not mentioned as part of an informal framework, since they are for “elementary” inferential statistics, they would be useful topics to consider in informal inferential reasoning.

Based on the existing frameworks, it appears that a few common trends emerge about informal inferential reasoning, and imply certain characteristics that good tasks would need.
Good tasks typically follow the PCAI model (Graham, 1987; Friel, O’Connor, & Manner, 2006), which highlights portions of a statistical investigation. Students should first pose the question(s), engage in collecting data, analyzing data, and then interpreting the results. Students will typically be given a contextual problem for which they have to formulate questions that can be answered by data. After they have formulated their questions, they must collect data, which could be from a randomized process such as simulations. Then, students would analyze the distribution of data using appropriate methods, and finally, they would draw conclusions or make inferences in the context of the problem by using the data as supporting evidence. This type of process supports informal inferential reasoning because the task is generally open-ended, and students would utilize the methods they know in order to justify their conclusions.

*Samples*

One of the main concepts needed to engage in informal inference is the idea of a sample. For the vast majority of problems we encounter in life, it is impractical to know every single unit of a population of interest. Because of this, we take one or more samples to learn information about the population without expending too much time or money. By taking samples from a population, we are collecting data that gives us information about the population. If samples are collected properly, then conclusions or predictions can be made about the population by using the sample as supporting evidence.

Most students begin learning about samples in elementary school, but they often have difficulty accepting its merits. According to Konold and Higgins (2003), students
do not want to generalize from samples because “(a) you can only know about the cases you observe [and] (b) to characterize a group, you must test every member of that group” (p. 196). Students appear to recognize that units in the population differ from one another, but they seem to have trouble accepting that one can make generalizations based upon the observed units if they are collected properly. For elementary school students who did accept making generalizations from a sample, they often favored biased sampling methods. For example, Konold and Higgins (2003) describe how students preferred self-selected samples, or techniques that emphasized “fairness” (p. 197). The reasoning behind this was to prevent feelings from being hurt, or to ensure that the individuals chosen would represent the population.

As students get older and become more experienced with collecting random samples, they begin to encounter other statistical ideas that should influence their conclusions. Several factors about the sample(s) need to be taken into consideration in order to make appropriate conclusions. These factors include: sample size, sampling variation, sampling distributions, and how the samples were collected. The following paragraphs will discuss these ideas along with common misconceptions associated with each.

Sample Size

Different authors have discussed two main beliefs that people tend to have about sample size prior to formal instruction. One belief that Kahneman, Slovic and Tversky (1982) discussed was “the law of small numbers,” which is the idea that any sample should represent the population accurately, regardless of its size. For example, in a sample of four
coin flips, most people would expect two flips to be heads, even though such a small number of flips could produce very unbalanced results. A variation of this belief is the notion that subsections of a long sequence of events will also accurately reflect the true proportions of the population, even though long strings of the same event often occur. If students believe that small samples are just as likely to represent the population accurately as large samples, then they may not think it is necessary to collect large amounts of data in order to make their inferences.

The second belief that people tend to have is the “empirical law of large numbers,” which is the intuition “that a large sample is better than a small sample for estimating a population parameter” (Sedlmeier & Gigerenzer, 1997, p. 35). For instance, many people would predict that a random sample of 100 adult males would give an average height closer to the true population mean than the average height of a random sample of 10 adult males. This belief would indicate that students may have an intuitive understanding of an appropriate way to collect data and draw conclusions, but various studies have suggested that students do not always apply the “empirical law of large numbers” even when it would be appropriate to take sample size into account.

Studies may have reached conflicting conclusions about the “empirical law of large numbers” based on the types of tasks or questions that participants were required to answer. Sedlmeier and Gigerenzer (1997) argue that studies yielded different results because the “empirical law of large numbers” works for some tasks, like frequency distributions, whereas it does not work for tasks that involve sampling distributions. Because different studies
generally utilized one of those types of tasks, they developed different conclusions. When participants answer frequency distribution tasks, they have to consider the results of one sample. The “empirical law of large numbers” tends to work for these tasks because participants only need to recognize that a larger sample is more likely to produce a statistic closer to the population parameter. In contrast, a sampling distribution task requires participants to consider a distribution of sample statistics from samples of a fixed size.

According to Sedlmeier and Gigerenzer (1997) “the empirical law of large numbers by itself is not sufficient to explain how sample size affects the variance of sampling distribution” (p. 44). In order to accurately answer sampling distribution questions, participants must consider how multiple sample statistics will be spread out, and how that spread will differ based on the fixed sample size that is taken.

The role of sample size is a topic that is easy to understand in some respects, while difficult to understand in others. Many students understand that a larger sample size is more likely to give statistics closer to the population parameter than a smaller sample size. However, when students have to consider the spread of multiple statistics taken from samples of the same size, it becomes much more challenging to make accurate comparisons and conclusions for large and small sample sizes. The successful completion of these tasks may also depend on the focus of the question. For example, studies have shown that attention to sample size tended to be more successful when participants were asked about the center of the sampling distribution as opposed to its tails (Well, Pollatsek, & Boyce, 1990). Thus, extreme values may be harder to predict based on sample size than measures of center.
**Sampling Variation**

One major type of sampling variation is variability *within* a sample. There are different ways to measure the variability within a sample, such as the range, the interquartile range, or the standard deviation. Of these three measurements, standard deviation is typically the most difficult one for students to grasp. According to Garfield and Ben-Zvi (2008), understanding variability and center is important because “when comparing groups or making inferences, we need to look at center and spread together: the signal, and the noise around the signal” (p. 204). In order to build a student’s conception of variability, they suggest starting with informal activities, like noticing that individual values vary, and working up to formal numerical measures of variability. Students should also recognize that variability can be due to both measurement error and an indication of the diversity of the subjects being measured. Garfield and Ben-Zvi (2008) describe a task that would allow students to discover the types of variability through repeated measures of head circumference and the measurements of different students’ heads (p. 210). This type of lesson would allow students to reason about variability informally first, so they will be better equipped to learn about specific measurements of variability, such as standard deviation.

Another type of sampling variation is variability *between* samples. This means that sample statistics often differ even when samples are taken of the same size, from the same population, and in the same manner. This occurs because the individual values collected for each sample are different between samples, so different values for statistical measures emerge. Since we generally make inferences on one sample, it is important to know how
samples can differ from one another, and use that knowledge to make sound judgments and express degrees of uncertainty. According to Wild, Pfannkuch, Regan, and Horton (2011), “any conceptual approach to statistical inference must flow from some essential understandings about the nature and behavior of sampling variation” (p. 253). Thus, it is imperative that students have a solid understanding of variation between samples, so they can consider which statistical measurement values are likely and unlikely to occur.

In the past it has been difficult to gauge students’ understanding of variation between samples because most test items asked for the theoretical value instead of a range of values. To remedy this lack of research, Watson and Kelly (2004) created a questionnaire that asked students to write a list of six sets that would give appropriate estimates for the number of times a spinner would land on the shaded part after 50 spins. If a student wrote the same number six times, then this showed a lack of appreciation for variation between samples. In contrast, if a student wrote six different numbers that encompassed more extreme values than one would expect, then this would indicate a lack of awareness about appropriate variation. Other questions on this questionnaire gave visual depictions of 27 sets of 50 spins for the same spinner. Two of these dot plots were made up; one distribution was perfectly symmetrical, while the other displayed an inappropriate amount of variation. According to Watson and Kelly (2004), only 36% of the students answered the pretest question about the inappropriate variation correctly, which indicates “more difficulty in appreciating too much rather than too little variation” (p. 136). After students received instruction, this percentage went up to 57%, but this was still the lowest percentage compared to the other two similar
test items. Students seemed to have an appreciation for “randomness” in the sense that a perfectly symmetrical distribution is unlikely to occur for such a small number of sets, but the belief that “anything can happen” is still ingrained in many students, even after instruction. This acceptance of unreasonable variation can make it difficult for students in informal inference tasks, because they might reach conclusions that allow for more variation than their data suggests.

Sample size plays a key role in variation between samples that is not very intuitive. While students may recognize that a larger sample size yields statistics closer to the true parameter, they have difficulty connecting this to a lower variance between sample statistics. According to Sedlmeier and Gigerenzer (1997), students perform better with frequency distribution problems because they are involved in real-life issues, “whereas the rule that the variability of a sampling distribution decreases with increasing sample size seems to have only few applications in ordinary life” (p. 46). In practice, we often make inferences on one sample, but it is necessary to consider how the statistics of different sample sizes vary in order to validate our conclusions.

*Sampling Variation with Respect to Expectation*

Some studies have looked specifically at how students coordinate their expectations of a value and their willingness to allow for potential variations. Expectation is understood as the mean or proportion expected based on the population of interest. Watson, Callingham, and Kelly (2007) describe a study of students ranging from third grade to ninth grade which observed their approaches in a series of tasks, and attempted to code their conceptual
understanding of expectation and variation, and the relationship between the two. The coding scheme Watson, Callingham, and Kelly (2007) developed is shown below, with a brief description of each level (p. 93).

Table 1: A Table Describing Watson et al.’s (2007) Coding Scheme for Variation and Expectation

<table>
<thead>
<tr>
<th>Name of Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1: <em>Idiosyncratic</em></td>
<td>Little or no appreciation of either expectation or variation</td>
</tr>
<tr>
<td>Level 2: <em>Informal</em></td>
<td>Primitive or single aspects of expectation and/or variation and no interaction of the two</td>
</tr>
<tr>
<td>Level 3: <em>Inconsistent</em></td>
<td>Acknowledgement of expectation and variation, often with support, but few links between them</td>
</tr>
<tr>
<td>Level 4: <em>Consistent</em></td>
<td>Appreciation of both expectation and variation with the beginning of acknowledged interaction between them.</td>
</tr>
<tr>
<td>Level 5: <em>Distributional</em></td>
<td>Established links between proportional expectation and variation in a single setting.</td>
</tr>
<tr>
<td>Level 6: <em>Comparative Distributional</em></td>
<td>Established links between expectation and variation in comparative settings with proportional reasoning.</td>
</tr>
</tbody>
</table>

In this study, ‘expectation’ refers to the expected mean or proportion for a specific sample or population. Students operating at the *Idiosyncratic* level did not seem focused on the mathematical parts of the task. For example, “students were likely to explain outcomes in terms of their favorite numbers, of the position of lollies in the container, or of the sizes of their hands” (Watson et al., 2007, p. 94). The *Informal* level indicated that students were
considering expectation or variation in the context of the problem, but not realizing that both were present. The *Inconsistent* level was reserved for students who answered the question with inconsistent responses. An example given of this was the response “Anything can happen” in conjunction with phrases like “50-50” (p. 99). They also tended to focus on single features of expectation or variation, like describing one thing as being more than something else, instead of providing a range of values. In contrast, students at the *Consistent* level “were likely to recognize the need for consistency in suggesting ranges in relation to data values and, when specifically asked for explanations of words associated with variation, generally provided satisfactory responses for all terms” (p. 101). The *Distributional* level was reserved for students who could describe the relationship between expectation and variation in a single data set without much prompting. Students often mentioned the shape, and how values vary around an expected center. The highest achievable level, *Comparative Distributional*, met the requirements of the previous level and students were able to compare the expectations and variations for two different samples. These levels are intended to be hierarchical, and Watson et al. (2007) suggest using these codes to determine where students are developmentally so that appropriate tasks can be used to help them progress to the next level.

*Sampling Distributions*

A theoretical sampling distribution of a statistic shows all the possible values a statistic can take when samples are randomly selected given a fixed sample size. Most researchers agree that theoretical sampling distributions are both important for formal inference topics, such as confidence intervals and hypothesis tests, and very challenging to
successfully teach (Aguinis & Branstetter, 2007; Gourgey, 2000; Wybraniec & Wilmoth, 1999; Zerbolio, 1989). There are numerous reasons why the topic of theoretical sampling distributions is so difficult to understand. Before learning about theoretical sampling distributions, students typically first learn about parent distributions and a distribution of a sample. In order to truly understand sampling distributions, students need to compare them to the parent distribution, and they need to distinguish between an empirical sampling distribution and a distribution of a sample. The language used for the latter two is very similar, which probably does not help students understand that an empirical sampling distribution focuses on the statistics of multiple samples instead of just one sample.

Theoretical sampling distributions are often difficult to teach because they combine several different statistical ideas that are all important in understanding its underlying structure. Not only do students need to know about the statistic of a sample, how the sample size affects that statistic, how statistics can vary, and how values can be distributed, they also need to be able to put all of that together, and compare it to both the parent population and theoretical sampling distributions for other fixed sample sizes. If students have not mastered all of the aforementioned ideas, then they will be missing a necessary piece of the puzzle required to understand this complex topic.

Research on student performance with sampling distributions has indicated two major findings: 1. Students have difficulty accurately predicting which sample size would be more likely to yield unrepresentative results, and 2. Students do not always recognize that a question is inquiring about a sampling distribution. Watson and Moritz (2000) described the
first finding in a study about undergraduates answering questions about a theoretical hospital. Only 20% of the 95 Stanford undergraduates correctly answered that a smaller sample would more often have more than 60% boys born on a given day. According to Watson and Moritz (2000), the reason the students had trouble with this question is because the questions are complex and “involve comparison of the tails of the sampling distribution of large and small samples” (p. 46). The hospital problem was also used in a study about students in grades 5 through 11. Fischbein and Schnarch (1997) argue that students focused more on the ratio for the hospitals (and thus believed that the number of days for the small and large hospital would be equal) instead of considering the relevancy of sample size.

The second major finding about sampling distribution research is that students often have difficulty determining whether a question is asking about a sampling distribution or a frequency distribution. Three pieces of evidence support this theory. According to Sedlmeier and Gigerenzer (1997), participants in various studies were able to “construct realistic frequency distributions,” they tended “to recall sampling distribution tasks as frequency distribution tasks,” and they tended “to construct identical distributions when asked to construct frequency or sampling distributions” (p. 44). Most students without formal instruction do not realize that sampling distributions tend to be bell-shaped and narrower than the parent distribution, so when they construct them they often make them flatter like a frequency distribution. Students typically do not have trouble constructing a frequency distribution, but it is much harder to visualize a sampling distribution if they have never observed one before.
The topic of sampling distributions is difficult to teach, and it takes time to teach it well. Often text books can illustrate basic concepts adequately, but many text books briefly describe a theoretical sampling distribution, and then discuss how they can be used without going into much depth. Many teachers recognize the deficiencies in the text-book based approach, so they often begin by doing a hands-on activity with an empirical sampling distribution in which students each take a sample and then plot their sample statistic on a dot plot. Wybraniec and Wilmoth (1999) utilized a hands-on activity with sampling blocks, and stated that students’ “comments during class discussion and interpretations on the exams indicated that they understood the reason these procedures are necessary for making inferences” (p. 80). While this approach helps to introduce the idea of how a sampling distribution is constructed and how one can make inferences based on the sampling distribution, it does not necessarily help students make broad generalizations about the effect of sample size on sampling distributions. To create multiple empirical sampling distributions by hand would require more time than most instructors are willing to sacrifice, so many choose to use technology for further explorations.

Sample Collection

While there are many ways to collect samples, not all of the techniques lend themselves to making valid statistical inferences. For example, convenience samples often produce biased results, and inferences made based on them would not be trustworthy. According to Rossman and Chance (1999), students should be required to make decisions about appropriate sampling either through random sampling or simulations when they make
their inferences (p. 302). Without recognizing the need for randomization when designing an experiment, students will not have a fundamental grasp on when it is appropriate for inferences to be made. Giving students the opportunity to collect data is a crucial part of the informal inference process because it serves as the foundation upon which the inferences are formed.

Simulations

One way to collect samples is through simulations. A simulation is letting one event represent something else in order to easily repeat the event to construct a data set; for example, if the odds of having a baby girl is 50%, then you could let a coin flip of “heads” represent having a girl, assuming a fair coin. The purpose of simulations is generally to make some predictions of what is reasonable or unreasonable based upon the results of several simulations. Simulations are random in the sense that the observer cannot predict with accuracy what sequence of events will occur. Batanero, Green, and Serrano (1998) described a study in which children were posed with two sequences (one random, and one non-random) of 150 coin flips; when asked which sequence was more likely to represent 150 coin flips, “most of the children chose the non-random sequence, and the perception of randomness did not improve with age” (p. 120). Some of the incorrect reasons that children gave were that the strings of heads or tails were too long, or the results were not exactly fifty percent for each. Children tend to have expectations about what should happen, and they may not accept the randomness of events.

In order to help students appreciate the idea of randomness, they should start by
observing physical simulations of an event. According to Rossman and Chance (1999), it is best for students to use “physical simulations to become comfortable with the idea of repeated samples,” (p. 299) before beginning the use of technology to perform simulations. Hands-on simulations can occur in a variety of ways, such as coin flips, tossing dice, sampling marbles, etc. Discussions about sampling with and without replacement, and the unpredictable sequences in the results can extend naturally from these types of simulations. By allowing students to experience simulations in a hands-on way first, they will be better equipped to understand what is being simulated in a computer environment.

Many different technologies exist that allow users to easily run simulations. In some cases, the results of the simulations are represented in tables or charts, while in other technologies the results are depicted visually. According to Ben-Zvi (2000), “technological tools support enhanced accessibility of many statistical conceptions by permitting the transformation of purely symbolic presentations into spatial-geometric ones, which are easier to grasp and build cognitive models on” (p. 142). Thus, while many students may not understand formal statistical notation, technology can allow them to see the illustration of a statistical topic. Some of these technologies, like applets, may have predetermined features that are designed to demonstrate specific statistical topics, such as sampling distributions or confidence intervals. These technologies may also focus on specific facets of the larger topic in order to draw the user’s attention to a specific idea. Rossman and Chance (1999) state that “technology can also free students from computational drudgery, allowing them to concentrate on exploring properties of the inferential procedures such as the effects of the
sample size or confidence level” (p. 299). By quickly allowing users to make comparisons between large and small samples, students can make generalizations more easily. Other technologies allow the user to design an experiment and run a particular number of simulations. Some of these technologies, like *Probability Explorer* (v. 2.01, Stohl, 2002) and *Fathom* (v.2.11, KeyPress Technologies), may be more educational in nature, while others, like *Excel* and the graphing calculator, may be designed for other practical uses.

*Probability Explorer*

*Probability Explorer* is an interactive computer program that allows users to design experiments and run simulations. This program is geared primarily towards an elementary and middle-school audience because of its ease of use, aesthetic appeal, and simple approach to exploring statistical topics. The screenshots below show what the program looks like when designing an experiment and running simulations.
Figure 1: Designing an Experiment in Probability Explorer
Drier (2001), the creator of *Probability Explorer*, states that “the overall goal in designing *Probability Explorer* was to create an open-ended environment for children to simulate interesting chance situations…they can design experiments of interest to them – whether they are playfully contrived or meant to model real-world phenomena” (p. 23). Users who are just beginning to use the program have the option to see the sample being collected, which can help reinforce the idea of what simulations are representing. One positive feature of the program is that the charts and graphs are dynamically linked to the simulation being performed, so students can see how the percentages or numbers are changing while the simulation is running. According to Drier (2001), this feature “allows
children to use graphs both as objects of display after a simulation is complete and as objects of analysis during a simulation” (p. 23). One of the drawbacks to this program is that while it can help students build the conception of a simulation, it does not perform simulations quickly or allow an easy storage of the results of several simulations for students who want to analyze the data further. Although there are limitations to what it is capable of doing, Probability Explorer is an excellent tool to help build informal inferential reasoning because it gives users the opportunity to collect data in a meaningful way, and then make judgments based on that data.

Fathom

Fathom is another interactive computer program that allows users to build statistical knowledge. While Probability Explorer is directed towards a younger audience, Fathom offers more formal statistical features that would be useful in a high school or introductory college statistics class. According to Meletiou-Mavrotheris (2003), “The designers of Fathom have drawn on constructivist theories of learning as well as several years of academic research about the way students learn and process statistical concepts and the main difficulties they face” (p. 270). Fathom allows users to design an experiment, run simulations, and collect the statistics from those simulations; this ability helps users to consider the spread of statistics for specific sample sizes. The screenshots below show various aspects of the program like running simulations, collecting measures, and generating dot plots.
Figure 3: Simulating Random Integers in *Fathom*
According to Meletiou-Mavrotheris (2003), the “Collect Measures” feature “can collect the statistic from repeated samples taken from the original population” (p. 273). This feature is useful because multiple samples can be run at the same time, and the statistics of interest can be collected and stored in an efficient manner. Based on these collected statistics, users can generate graphics like dot plots. This feature gives students the opportunity to see an empirical sampling distribution based on their data. Although Fathom visually demonstrates sample-to-sample differences that would help build informal inferential reasoning, the collection process is not necessarily intuitive and does require time and effort to learn.

Figure 4: Collecting Measures and Generating Dot Plots in Fathom
Some technologies that run simulations were primarily designed for other uses as well. For example, *Microsoft Excel* (2007) is a computer program designed to organize data in a spreadsheet, but it can also compute simple statistical measures and run simulations. Users can randomly generate numbers by using the “randbetween( )” feature, and can tally the value of their choice by using the “countif( )” command. Screenshots of these commands in *Excel* are shown below.

![Figure 5: Generating Random Numbers in Excel](image)
According to Drier (2001), users can use the aforementioned commands in conjunction with the F9 key to “execute a series of random events, tally and display the numerical results, and change a graphical display of the experimental results” (p. 171). The F9 key performs a new simulation almost instantaneously, but does not save the old results. It is possible to take multiple samples of the same size in Excel, but to do so would require formulas to be populated in a vast array, and if the sample size is large (or if many samples were being taken), then the user would be required to scroll through the spreadsheet to observe all of the information. It is also possible to create a histogram linked to the sample means. According to Moen and Powell (2005), the FREQUENCY function in Excel can
update the diagram chart because “when the sample means change (every time function key F9 to recalculate is depressed), the frequency counts will automatically be updated, which in turn will update the histogram” (p. 37). While it is conceivable to simulate multiple samples at the same time in Excel, it is not intuitive or easy to learn for beginners. In spite of its complicated formulas, Excel has the advantage of being ubiquitous, whereas Fathom and Probability Explorer are not as readily available for teacher use.

**Graphing Calculator**

Another tool that is widespread in most mathematics classrooms is the graphing calculator, which also has the ability to run simulations. Users can run simulations on the graphing calculator by using the “randINT( )” function, and the values that are given can be stored in a list. Using the “randINT( )” function on the graphing calculator is shown in the screenshot below.

![Generating Random Integers in a Graphing Calculator](image)

Mulekar and Siegel (2009) describe how a class could use a TI-83 or TI-84 graphing calculator to simulate a sample, obtain basic statistics on that sample, record the sample
mean, and repeat the process in order to have a collection of sample means. Although it is possible to build a collection of statistics, this would be very time-consuming, and the graphics offered on the calculator do not include dot plots, so the empirical sampling distribution would have to be viewed as a histogram. Additionally, if all students in a classroom are running simulations for the same problem, the teacher would need to tell the students to choose different seed values so the simulations will generate different results. It is not impossible to create an activity with the graphing calculator to build empirical sampling distributions, but aside from its ubiquity, it does not offer many advantages over other technologies.

_The Teacher’s Role_

Teachers have the difficult task of not only understanding the statistical content previously mentioned, but also using that knowledge to help guide their students in the informal inferential process. The first requirement of content knowledge is challenging by itself because mathematics teachers do not always have strong backgrounds in statistical topics. Many mathematics education programs do not place much emphasis on statistics classes or statistical topics, so teachers may not have much statistical knowledge to teach these topics. Without the requisite content knowledge, teachers will struggle to teach these topics successfully, thereby transferring their own misconceptions to their students. Undergraduate students and pre-service teachers have had difficulties understanding basic statistical topics like the mean (Leavy & O’Loughlin, 2006; Mevarech, 1983), the median (Groth & Bergner, 2006), and thinking about distributions (Leavy, 2004, 2006).
Additionally, while teachers may know how to calculate common statistics like the mean or standard deviation, they do not necessarily recognize these as tools to be used in data analysis. According to Makar and Rubin (2009), “it is vital that the focus in using statistical tools is embedded in the reason that we do statistics – to understand underlying phenomena” (p. 84). In order to teach informal inferential reasoning successfully, teachers need to realize that statistical tools can help students make inferences, but they should not be the culminating result.

Once teachers have the required content knowledge, they need to be able to transform that to pedagogical content knowledge. Teachers need to be able to choose open-ended tasks that are appropriate for their students, and they need to know appropriate questions to ask that will keep students focused on making inferences by supporting them with data. Leavy (2010), described a study of 26 Irish pre-service teachers who worked in groups of five or six to design a lesson study that focused on a statistical activity which would incorporate major components of statistical inference. From this study, two observations were made: teachers had difficulty using effective questioning to build informal inference, and teachers did not always know how to handle unanticipated responses from children. Because of this, Leavy (2010) suggests that “there was a failure to focus children on analysis of the data, the identification of patterns, and the generation of assertions arising from those patterns” (p. 57). In terms of questioning, the teachers did not recognize the importance of using data as evidence for the inferences being made, so they did not ask students to justify their conclusions using data. The unanticipated responses from children, such as conjecturing why
one result would be better than another before considering the data, caught teachers off-guard and redirected the focus of the lesson away from the data. This suggests some challenges that teachers face when giving open-ended tasks to their students. Teachers must have good questions in mind to redirect students’ attention to the problem, and they need to be prepared for all types of responses from students.

In order to effectively teach some of the higher level informal inference tasks, teachers also need to have knowledge about the technologies they are planning on using in their classrooms. According to NCTM (2007), “if teachers are to learn how to create a positive environment that promotes collaborative problem solving, incorporates technology in a meaningful way, invites intellectual exploration, and supports student thinking, they themselves must experience learning in such an environment” (p. 119). Many different technologies exist that can be used for different types of tasks, with some illuminating concepts better than others. Teachers have to make a decision about which technology to use based on what is available, what will help students learn the topic best, and what students would need to learn about the technology itself.

Research Questions

This study considers fifty-seven pre-service and in-service teachers in three institutions across the United States who engaged in an informal inference task. Before solving the task, the participants were allowed to choose from one of four technologies: Probability Explorer, Fathom, Microsoft Excel, and the graphing calculator. The task, which will be described in more detail in Chapter 3, focused attention on reasonable vs.
unreasonable values, sample size, and using data to justify conclusions. The author chose the following research questions to focus on:

1. How do teachers use different technology tools to conduct simulations and make informal inferences?

2. How do teachers reason about variation with respect to expectation and sample size, and how do they use data to develop reasonable predictions?
A multiple case study methodology was used to analyze this data. According to Baxter and Jack (2008), “a multiple case study enables the researcher to explore differences within and between cases” (p. 548). This observational study utilized written documents by teachers as they used a technology of their choice to solve an informal inference problem. The cases were defined by three of the technologies teachers could use in a simulation task, with each case having multiple teachers’ responses. The purpose of choosing technology type to determine the cases was to explore the similarities and differences between how teachers formed their conclusions and how their technology choice may have influenced those conclusions. This chapter seeks to address the context of the study, the data used, and the analysis techniques employed to examine the cases.

**Larger Study**

The data for this study is a part of a larger study that considered how teachers utilized dynamic statistical software to solve three exploratory data analysis tasks. According to Lee, Kersaint, Harper, Driskell, and Leatham (2012), the “research group examined teachers’ use of dynamic statistical technology environments with teachers enrolled in courses from eight different institutions in the United States in which faculty were using the same curriculum materials” (p. 291). The curriculum materials came from *Preparing to Teach Mathematics with Technology: An Integrated Approach to Data Analysis and Probability* (Lee, Hollebrands, Wilson, 2010). Data from six tasks were collected across institutions.
Instructors at each institution were able to choose whether their teachers participated in each task.

**Context for the Study**

The task examined in this study came from Chapter 6 (the final chapter) in the Lee et al. (2010) book. Task responses were collected from 62 pre-service and in-service teachers across three different institutions in the United States. The teachers were enrolled in classes that focused either on using technology to teach middle or high school mathematics, or teaching statistics in elementary or middle school classes. Most of the teachers enrolled in these classes were pre-service teachers, but a few were in-service teachers or graduate students. The informal inference task that the teachers completed required them to choose one of four different technologies to simulate data, and then they were expected to make an inference and justify their conclusions based on data. In Chapter 6, teachers were taught how to simulate data for similar tasks using *Probability Explorer*, and *Fathom*. In the previous chapter, teachers learned how to use *Microsoft Excel* and the graphing calculator to run simulations. While solving the task, teachers recorded their responses in word documents that also included screenshots to help justify their conclusions. Some of the documents were very short (less than a page including screenshots), while others were extremely detailed (fifteen pages). The task that teachers solved is described below.
The Task

A fast food restaurant is giving away prizes with each value meal. When each value meal is purchased, the customer gets to randomly choose one of four cards. Each card displays a different gift: coffee mug, bumper sticker, hat, ink pen. The hat is the most expensive item for the restaurant to give away. Thus, the manager is particularly interested in the amount of hats that may be given away.

a) Use a technology tool (graphing calculator, Excel, Probability Explorer, or Fathom) to simulate this context and examine results from simulated data.
   i) Justify your choice of technology. Explicitly describe how you modeled the context and simulated the “give away” with the technology tool.
   ii) If the restaurant runs this special “give away” for a weekend (Friday through Sunday), what are some reasonable expectations for the proportion of customers that will receive hats?

b) Prepare a response to the following question. Include any screenshots of how you used the simulation you created in part (a) to investigate the question:

On Friday, 93 value meals were purchased. On Saturday, there were 387 value meals purchased. Since the hats are the most expensive item, the manager needs to carefully check the records for each day to be sure a reasonable number of hats were given away. For each day, what number of hats given away should seem unusual or surprise the manager? Justify your reasoning.

Figure 8: The Informal Inference Task Teachers Solved

The four technology types teachers could choose from were Fathom, Probability Explorer, Microsoft Excel, and the graphing calculator. Four teachers chose to use Fathom, thirty-one chose Probability Explorer, twenty-two chose Microsoft Excel, and five chose to use graphing calculators. I decided to exclude the five graphing calculator responses from my analysis because a cursory read through the responses showed that two of the teachers chose arbitrary numbers for their conclusions, and another two did not answer the question entirely. Based on this, I decided not to include these responses because it was not evident how the use of the graphing calculator informed the conclusions the teachers made. In contrast,
although only four teachers used Fathom, I chose to include those responses because it was clear how the teachers used that technology tool to solve the problem. After excluding the graphing calculator responses, this left a total of 57 responses, grouped among three different cases.

**Analysis Techniques**

Before I read any teachers’ responses, I attempted to solve the task using the different types of technologies to get a sense for how one might go about solving such a task informally. Then I tried to pinpoint the technological and statistical skills that were needed to solve the task. To gain an awareness of how teachers solved the task, I read a few of the responses to see how they were structured.

Responses were separated into the three cases. I read each response and made general notes about the approaches used. Based on the nature of the task and on the review of literature, there were certain aspects I looked for actively (e.g. teachers’ attention to variation with respect to expectation and sample size, repeated sampling). Other themes emerged that were also interesting to me, so I made notes of those as well. Upon identifying these items of interest, I developed an open coding scheme that concisely recorded certain features of each response. According to Moghaddam (2006), a researcher who uses an open coding scheme “engages in breaking down, analyzing, comparing, labeling, and categorizing data” (p. 56). The purpose of using this coding scheme was to read through the responses and find general themes across the selection. Shown below is a simplified organization that shows the general problem-solving strategies teachers took, as well as the relevant questions that needed to be
answered about each response. The bulleted questions below the diagram offer a preview for what my coding schemes will focus upon.

![Figure 9: An Organization of the Problem-Solving Pattern and Related Questions](image)

**Coding Initial Expectations for the Proportion**

The second portion of part (a) in the informal inference task (Figure 8) requires teachers to determine reasonable expectations for the proportion of customers who will receive hats over the weekend. This prompt does not give an associated sample size, so teachers have to determine whether to provide a range of values, and if they do, what sample size to base that range on. In order to code the initial expectations for the proportion, a modified version of the framework described in the literature review by Watson, Callingham, and Kelly (2007) was used. This modified framework is described below.
Table 2: Coding Scheme for Initial Expectations for the Proportion

<table>
<thead>
<tr>
<th>Inconsistent – 0</th>
<th>Pre-Consistent – 1</th>
<th>Consistent – 2</th>
<th>Distributional – 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Not recognizing appropriate variation of expectation</td>
<td>• Listing single value based on one sample (of any size)</td>
<td>• Appreciation of variation around a reasonable center (but too much or too little)</td>
<td>• Appropriate range around 25% based on sample(s) OR</td>
</tr>
<tr>
<td>• Too much variation: Belief that “anything can happen”</td>
<td>• Recognizing some variation: “about 25%” but without an associated range</td>
<td>• A reasonable range, but no indication of the sample size or data used to construct</td>
<td>• Based on thoughts of how the range would differ based on different sample sizes</td>
</tr>
<tr>
<td>• No variation: Point estimate</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Some of the levels described by Watson et al. (2007) were not included in this coding scheme because they were either too simplistic for this group of teachers, or they were too complex for the initial question. Additionally, the levels that were utilized were modified to match the details of this task. Each of these levels is described below.

Teachers operating at the Inconsistent level (0) either had an expectation of far too much variation for the proportion, or did not appreciate any variation and believed that the expected proportion would be exactly 0.25. Teachers functioning at the Pre-Consistent level (1) either simulated one sample of any size, and listed a single value based on that sample, or they stated that the expected proportion would be “about 25%.” This level was not in the original framework, but it was added out of necessity to bridge the gap between the Inconsistent and Consistent levels. The Consistent level (2) was for teachers who provided a range around a reasonable center (a proportion equal to or close to 0.25) but with too much or too little variation based on their simulations. Teachers could also be coded at the Consistent
level if the range they provided was reasonable, but there was no indication of the sample size or data used to construct it. This would be coded Consistent instead of Distributional because the range provided may not be appropriate depending on the sample size for the weekend. The Distributional level (3) was the highest possible code for the initial expectations. Teachers who were coded at this level either provided an appropriate range based on sample(s) of a specified size, or they made statements comparing the ranges for hypothetical sample sizes. For example, if a teacher said that a larger sample size would produce a smaller range of proportions around 0.25, then that response would be coded as Distributional.

Coding Variation with Respect to Sample Size

Part b) of this task requires that teachers attend to sample size in order to approach it in an appropriate way. Because teachers have to consider numbers or proportions that would be deemed unusual for two different sample sizes, they should consider how the proportions vary based on sample size. Many teachers explicitly stated that the larger sample size would have sample proportions that varied less than those of the smaller sample size. However, the task does not directly ask the teacher to comment on how sample size affects the variability so many teachers did not mention it. To code teachers’ appreciation of variation with respect to sample size, the previous framework was modified to include comparisons between Friday and Saturday. The organization for this is shown below.
Table 3: Coding Scheme for Variation with Respect to Sample Size

<table>
<thead>
<tr>
<th>Inconsistent – 0</th>
<th>Distributional – 1</th>
<th>Pre-Comparative Distributional – 2</th>
<th>Comparative Distributional – 3</th>
<th>Explanatory Comparative – 4</th>
</tr>
</thead>
</table>
| • No appreciation for appropriate variation  
  • “Anything can happen” OR  
  • Exactly 25% for each sample size | • Same proportion ranges for Friday and Saturday (no appreciation for sample size)  
  • Weak appreciation for two aspects, but not for the third (aspects: variation, expectation, sample size) | • Ranges based on data but given too much extra “wiggle room”  
  • Inappropriate variation with respect to sample size, but based on empirical evidence  
  • Fairly reasonable range(s), but no indication of the data used to construct it | • Reasonable ranges with appropriate variation provided for the different sample sizes | • Meets the requirements of Level 3 AND  
  • Explicitly states that a larger sample size would produce values more likely to be closer to the expectation (less variation) |

Teachers operating at the Inconsistent level (0) were very similar to those for the initial expectation in the sense that there was no appreciation for appropriate variation. For this coding scheme, the Pre-Consistent and Consistent levels were left out because teachers either operated at an Inconsistent level, or they operated at a Distributional level that did not make appropriate comparisons for different sample sizes. In this instance, teachers functioning at the Distributional level (1) typically appreciated two of the aspects of the problem while ignoring the third aspect. For example, a few teachers gave the same proportion ranges for both Friday and Saturday, and while this could show an appreciation for expectation and variation, it does not consider how proportions vary with respect to
sample size. Because they missed this crucial piece of the problem, they would be coded at the Distributional level.

The Pre-Comparative Distributional level (2) was added to transition between the Distributional and Comparative Distributional levels. Teachers were coded as Pre-Comparative Distributional when they provided ranges that were based on data, but they widened the range to include proportions that had a very low probability of occurring. These ranges tended to be narrower proportionally for the higher sample size, but they still encompassed an inappropriate amount of variation. A teacher could also be coded at this level if the ranges were based entirely on empirical evidence, but the variation for the higher sample size was larger. While the ranges were based on data, the teacher could have recognized that the extreme values rarely came up for the larger sample size, so s/he is showing a lack of observation about typical variations. The final type of response that was coded as Pre-Comparative Distributional occurred when a teacher gave a range or ranges that were fairly reasonable, but did not indicate where the numbers came from or how they were found.

Teachers functioning at the Comparative Distributional level (3) gave reasonable ranges for both Friday and Saturday with appropriate variation provided for the different sample sizes. According to Watson et al. (2007), learners operating at the Comparative Distributional level “required an understanding of the sample mean as a representation of the population mean in the two sample sizes, as well as the explanation of multiple aspects of potential variation associated with the values in the two samples” (p. 108). Thus, teachers
operating at this level have an understanding of expected proportions, variation around those proportions, and how that variation is different for different sample sizes. Since this task did not specifically ask teachers to explain the differences in variation, the Comparative Distributional level for this particular task does not require an explanation about the variation. However, since so many teachers did elaborate about variation in their responses, a fourth level was created named Explanatory Comparative Distributional level (4). Teachers operating at this level meet the requirements of the previous level, and explain that larger sample sizes tend to have less variation of sample proportions while smaller sample sizes tend to have more variation of sample proportions.

Coding Repeated Sampling

Another embedded statistical topic in the task was repeated sampling. In order to determine the unusualness of the number of hats, teachers had to take a number of samples to determine what was considered usual. During the initial reading of each document, I made notes about the number of samples the teacher took for sample size 93 and sample size 387. The reason I focused on those sample sizes is because I was mainly concerned with the comparisons they would make between the sample sizes, so I chose to focus only on the aspects of the responses that dealt with those sample sizes. Some teachers only took one sample of each size, while others took 100 of each size. Because there was such a wide variety of the number of samples taken, I made categories that indicated a range of samples that teachers took. The numeric categories I chose were 0 samples taken, 1 sample taken, 2 to 4 samples taken, 5 to 9 samples taken, 10 to 20 samples taken, 21 to 50 samples taken, and
51 or more samples taken. These categories do not have the same numeric width because I interpreted a larger difference in how students approached the problem for lower numbers vs. higher numbers. For example, I think a teacher who took two samples compared to a teacher who took five samples should be in different categories because the teacher who took more samples is showing an awareness that more samples will give more information. However, I do not think that teachers who did 100 repeated samples vs. 103 repeated samples should be placed in different categories because a difference of three in such a large number of samples is negligible. Based on this premise, I chose to make the ranges for the beginning intervals smaller, while they gradually became larger to encompass more types of responses.

An issue that came up while coding each response was the use of language eight teachers utilized. Instead of specifying the exact number of samples they took for each day, they used words like “several,” “many,” “numerous,” etc. In order to analyze these responses with the rest of them, I placed them in the 5 to 9 category. It is my supposition that when we use words like several or many, we mean approximately more than two, and likely a relatively large amount. But without specification, it is reasonable to assign these terms to the category of less than 10.

Another concern that came up while coding is that teachers did not always explain exactly what they did when stating their conclusions. Some teachers gave no indication of how many samples they took, and in those cases I did not feel comfortable placing them in a category. Because of this, there is one category for these responses, coded “Unclear,” in terms of repeated sampling.
After coding the responses to be in a numeric category, it was evident that some of the categories had many more responses than others, and it was difficult to notice trends when the numbers for the categories were so unbalanced. To remedy this, I decided to create three general categories for types of samplers: low, medium, and high. Teachers who took 0 to 4 samples were considered “low samplers,” 5 to 20 samples taken were “medium samplers,” and 21 or more samples taken were “high samplers.” Again, the numeric widths are not the same for the reasons previously mentioned. By narrowing it down to three categories, the numbers became more comparable and trends were easier to notice. I am also more comfortable doing this because I feel confident that the teachers who wrote about taking “several” or “multiple” samples should be included in the “medium samplers” category.

*Formation of Conclusions*

While some of the aforementioned coding schemes are based on existing frameworks, the next trends that I coded were based on my observations of emerging patterns while reading the documents. The trend I noted was how teachers formed their conclusions, and whether they were stated directly or implied by the screenshots of their technology use. It became apparent that since different types of technologies utilize different visual aids, teachers might become focused on certain aspects of the data depending on the technology they chose. In the vein of a grounded theory approach, I decided to include this trend in my analysis even though I had not considered it previously. According to Cousin (2005), this approach “…requires that, in the first instance, researchers try to see what the data are telling
them rather than asking those data to yield responses required by the issues or hypothesis that guided their collection.” (p. 425). Thus, my approach to this study included both propositions for analysis, as well as a pseudo-grounded theory approach. I decided upon seven main categories for forming conclusions, with an eighth category for conclusions that were unclear. The categories are listed and briefly described in the table below.

Table 4: Types of Conclusions Formed with Descriptions

<table>
<thead>
<tr>
<th>Type of Conclusion</th>
<th>Brief Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupported by Data</td>
<td>Conclusions were not supported by data</td>
</tr>
<tr>
<td>Extremes for Any Outcome</td>
<td>Ranges built using extreme numbers for any outcome</td>
</tr>
<tr>
<td>Extremes for Specified Outcome</td>
<td>Ranges built using extreme numbers for specified outcome</td>
</tr>
<tr>
<td>Excluding Extreme Values</td>
<td>Ranges built excluding observed extreme values</td>
</tr>
<tr>
<td>Formal Procedure</td>
<td>Formal statistical procedure used to build ranges</td>
</tr>
<tr>
<td>Informal Procedure</td>
<td>Informal procedure used to build ranges</td>
</tr>
<tr>
<td>Multiple Strategies</td>
<td>More than one strategy used to build ranges</td>
</tr>
<tr>
<td>Unclear Strategy</td>
<td>Strategy used was not clear</td>
</tr>
</tbody>
</table>

A teacher’s response was coded as “Unsupported by Data” if the ranges given for Friday and Saturday did not appear to be based on data. For example, one teacher wrote “The owner should be concerned if the total number of hats given away is well above 25% of the total number of value meals, so 30-35% should be alarming.” This would be coded as “Unsupported by Data” because the teacher did not use data to support this assumption, and
s/he made no distinction between alarming percentages for Friday and Saturday. While the assumption may not be invalid, it is not supported by evidence because the teacher does not explain how those numbers were found, and screenshots do not support the statement.

Another way that conclusions were formed was by considering the values for all prizes, and choosing the most extreme values to build the range. Some teachers explicitly stated doing this, while others implied it based on their screenshots. For example, one teacher stated “Since the hat has the equal chance of occurring as the other gifts, we must also look at the other gift data.” In addition to that statement, the teacher also showed a screenshot that displayed extreme numbers for different prizes, and then the range given for Friday matched those extreme numbers. Based on both pieces of information, this response would be coded as “Extremes for Any Outcome,” because the teacher is considering extreme numbers regardless of the type of prize. Other responses were coded as “Extremes for Any Outcome” based only on the screenshots. Some teachers did not clearly state how they formed their conclusions, but the supporting evidence they included indicated that they looked at any prize to form the lower and upper bounds of the reasonable ranges. For instance, one teacher who used Probability Explorer found that an acceptable range of hats given out on Friday could be 18 hats to 26 hats. This teacher took three samples of size 93, and two of those simulations are shown below.
At the beginning of the teacher’s response, s/he stated that the number 3 represented the hat. However, in the three samples that were simulated, the hat came up 21, 23, and 24 times. Interestingly, these numbers were not used to build the range. Based on the screenshots, it is my supposition that the teacher used the lowest (18) and the highest (26) values for any prize to build the range. This type of response would be coded as “Extremes for Any Outcome.”

Another common way to form conclusions was by using only the extreme numbers that were simulated for hats. Some teachers utilized methods that were only concerned about hats, so it was easier to determine that they focused on the extreme values for hats. For example, teachers who used Excel often counted only the random numbers that represented hats, so they did not focus on the other prizes. Other teachers included screenshots that showed the percentages for each type of prize, but then the numbers for the ranges given corresponded with the numbers for hats. For example, one teacher included the following information in their response:

![Figure 10: Two Data Tables of the Simulated Results of Two Samples in Probability Explorer](image-url)
The teacher then states that “On Saturday, up to approximately 118 hats or as few as 85 of hats can be given away.” While 118 is the high extreme for any of the prizes, 85 is not the low extreme of all of the prizes. Based on the teacher’s given range and the screenshot, it appears that s/he is only looking at the extreme values of hats to build the range. Therefore, this response would be coded as “Extremes for Specified Outcome” because the teacher is only considering the data for the particular prize of hats.

Another popular option for forming conclusions was by excluding the most extreme values to build the ranges. Again, some stated outright that they were excluding the most extreme, while others included screenshots that implied they were excluding extreme values. For example, one teacher stated “I would suggest to the manager that for Saturday, he should not expect to give out more than 110 hats, because I only got 2 out of about 40 trials that were above 110.” This response would be coded as “Excluding Extreme Values” because the teacher is indicating that s/he did observe extreme numbers, but since they did not occur.

Figure 11: Two Tables of the Simulated Results of Two Samples (size 387) in Excel
often, decided not to include them to build the ranges. Other teachers used screenshots to help support their conclusions, and made it easier to determine that they had in fact excluded the most extreme values. For example, one teacher included the following in the response:

![Dot Plot in Fathom](image)

The teacher then states that an acceptable range for Friday would be between 16 and 32 hats. Based on the screenshot, you can observe that one value is below 16 hats and one value is above 32 hats. Because of the range given in comparison to the screenshot, this response would be coded as “Excluding Extreme Values”

A few teachers did not fit into any of the previous categories, but they did write about their techniques for solving the problem. The following categories did not have many responses, but to be true to the teachers’ descriptions I decided to include them. The first
category is “Formal Procedure.” Responses were coded “Formal Procedure” if teachers used formal statistical procedures to solve the problem. For example, one teacher performed Chi-Square tests in *Microsoft Excel* to determine whether individual values would be considered unusual. Another teacher was coded as “Formal Procedure” because s/he ran ten simulations and then found the mean and standard deviation to help determine a range of acceptable values. The screenshot used for that response is shown below.

![Figure 13: A Spreadsheet in Excel Showcasing Formal Procedures](image)

Responses were coded as “Informal Procedure” if the mathematical procedure or organization used was very simplistic and/or did not make sense. For example, one teacher ran two simulations each for Friday and Saturday, and then added the low values to determine a lower bound, and the high values to determine an upper bound. Another response that was coded as “Informal Procedure” did not actually produce any ranges, but focused on whether or not a simulation produced percentages above or below 25% instead of
considering how far those percentages deviated from the expected value. A few responses were coded as “Multiple Strategies” because a teacher utilized a combination of the strategies. For example, one teacher focused on the extreme values for hats on Friday, but excluded the most extreme values of hats for Saturday. Since two different strategies were employed for Friday and Saturday, this was coded as “Multiple Strategies.”

Other responses were not placed in any of the aforementioned categories because it was not clear what approach the teacher decided to take. These occurred for a few different reasons. First, some teachers did not actually state any ranges for Friday or Saturday, so none of the previous categories would make sense for those teachers. Second, some teachers determined reasonable ranges for Friday and Saturday, but they were vague in their descriptions and/or their screenshots did not match up with their ranges. Finally, some teachers indicated that they chose extreme values to build the ranges, but it was not obvious whether they were choosing an extreme number based only on hats, or an extreme number based on any prize. Because of this lack of clarity, responses that met the previous criteria were coded as “Unclear Strategy.”

Looking Across Cases

In order to look across cases and see general trends that were occurring that were not based on technology choice, I used the previous codes for types of sampler and looked at the trends that occurred based on the number of repeated samples simulated. Teachers were coded as a Low, Medium, or High sampler, and four teachers were coded as Unclear because it was not clear how many samples they simulated. By looking across the cases based on the
type of sampler a teacher was considered, it was easier to determine how the type of sampler was related to the types of conclusions. Additionally, the number of samples simulated also appeared to be related to the reasoning levels for part (b) of the task, so a cross-analysis of the two variables was done as well.

Furthermore, because teachers focused different levels of detail on parts (a) and (b) of the task, it seemed necessary to consider how their levels of reasoning matched up from one part of the task to the next to observe any discrepancies or lingering misconceptions. The Inconsistent levels matched up well for both parts of the task. The Pre-Consistent level had two components; if teachers listed a single value, then they would probably operate at the Inconsistent level for part (b) of the task, but if they stated “about 25%” then they might be at the Distributional level in part (b) by attending to two aspects. The Consistent level matches with Pre-Comparative (if the teacher gave too much variation) or the Distributional level (if the teacher did not specify the sample size used). If teachers were at the Distributional level for part (a), then they should be at the highest levels. By comparing levels from part (a) to part (b), it is possible to determine which part teachers emphasized most.
CHAPTER 4
RESULTS BY TECHNOLOGY CASE

The purpose of this chapter is to describe the teachers’ work on the task for each of the different cases of technology type. This chapter will be organized into three major sections examining responses for Probability Explorer, Excel, and Fathom. Each section will include a general description of the documents as well as the most popular justifications for that choice of technology. The sections will also indicate the levels for the initial expectations as well as the final expectations for the proportion of hats for each technology case. Finally, trends in repeated sampling and the formation of conclusions will also be shown for each technology case. A fourth section will highlight the main similarities and differences between the technology cases.

Probability Explorer

Thirty-one teachers chose Probability Explorer to solve the restaurant task, making it the most popular technology choice. The documents the teachers submitted varied widely in length, ranging from short (1 page) to long (12 pages). A typical response was generally between 2 to 5 pages in length. Some teachers were incredibly detailed and described exactly what they did, while other teachers wrote murky details, or allowed their screenshots to speak for themselves. Six teachers included a large number of screenshots (10 or more), but the screenshots typically showed the results of individual samples to be used as supporting evidence. Thirty teachers showed a screenshot of at least one data table in their documents, while sixteen showed pie charts, and thirteen showed bar charts.
Reasons for Choosing Probability Explorer

The first part of the task asked teachers to justify their choice in technology. While five teachers did not answer this part of the task, the rest of the Probability Explorer users did. The following table shows the most common reasons for using Probability Explorer, along with the percentages of teachers who specified each reason.

Table 5: Top Justifications for Choosing Probability Explorer

<table>
<thead>
<tr>
<th>Justification</th>
<th>Frequency</th>
<th>Percentage (out of 31)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic Nature</td>
<td>16</td>
<td>52%</td>
</tr>
<tr>
<td>Ease of Use</td>
<td>14</td>
<td>45%</td>
</tr>
<tr>
<td>Available Representations</td>
<td>12</td>
<td>39%</td>
</tr>
<tr>
<td>Fun/Entertaining</td>
<td>6</td>
<td>19%</td>
</tr>
<tr>
<td>Speed</td>
<td>4</td>
<td>13%</td>
</tr>
</tbody>
</table>

The most popular justification teachers gave for choosing Probability Explorer was its ability to dynamically link the simulations with the graphical representations, and to visualize the simulations as they were occurring. Teachers appreciated that they could see the changes as they were happening, and they knew what each symbol represented in terms of prizes. The second most popular reason for choosing this technology was its ease of use; many teachers described how simple it was to design an experiment and run simulations for
the appropriate sample sizes. Twelve teachers stated that the available representations (data
table, pie graph, bar graph) were another advantage in this technology choice because these
features were automatically available. Finally, six teachers said that the program was fun and
entertaining to use, while four teachers stated that the speed of the simulations was an
advantage. Most teachers who answered this part of the task gave multiple reasons, so they
found several advantages to using *Probability Explorer*.

*Initial Expectations for the Proportion of Hats*

In part (a) of the restaurant task, teachers are required to give reasonable expectations
for the proportion of customers who will receive hats in a weekend. This question does not
give an associated sample size, so it was left to the teachers to determine what would be
considered “reasonable.” Each response was coded as either Inconsistent (0), Pre-Consistent
(1), Consistent (2), or Distributional (3), as described in the previous chapter. The frequency
table below shows how many *Probability Explorer* teachers were coded for each category.
Four teachers did not answer this part of the task, so they are excluded from the table.
Most of the teachers who answered this question were operating at the Distributional level, meaning that they gave an appropriate range around 25% grounded by specific sample sizes, or based on thoughts that described how ranges would differ based on sample size. Six teachers functioned at the Consistent level, implying that their ranges were somewhat reasonable but gave too much or too little variation, or they may have been reasonable for specific sample sizes, but they did not include any data to support their conclusions. Teachers were most commonly coded as Consistent for the latter reason. Five teachers were working at the Pre-Consistent level, indicating that they either provided a single value for the expectation based on a sample, or they stated that the proportion would be “around 0.25”, or “roughly ¼.” Four teachers at this level stated they believed the proportion would be approximately 0.25. Finally, one teacher operated at the Inconsistent level because s/he gave a range based on one sample (size 500) and stated that reasonable values would be between 0.25 and 0.5. This showed a lack of appreciation for expectation (it was not centered around
and it did not recognize appropriate variation for the expectation. The upper bound of 0.50 was far too high for the specified sample size of 500.

**Reasoning About Variation with Respect to Sample Size**

Part (b) of the task asked teachers to consider what number of hats given away would be considered unusual for two different sample sizes (93 and 387). Many teachers gave a range of values or proportions for each day that would be considered reasonable, and then stated that anything outside those ranges would be considered unusual. Each response was coded as Inconsistent (0), Distributional (1), Pre-Comparative Distributional (2), Comparative Distributional (3), or Explanatory Comparative (4), as described in the previous chapter. Three teachers who used *Probability Explorer* did not answer this portion of the task, so they are excluded from the frequency table below.

Table 7: The Reasoning Levels for *Probability Explorer* Users in Part (b)

<table>
<thead>
<tr>
<th>Level</th>
<th>Frequency</th>
<th>Percent (out of 28)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inconsistent – 0</td>
<td>3</td>
<td>10.71</td>
</tr>
<tr>
<td>Distributional – 1</td>
<td>4</td>
<td>14.29</td>
</tr>
<tr>
<td>Pre-Comparative – 2</td>
<td>6</td>
<td>21.43</td>
</tr>
<tr>
<td>Comparative – 3</td>
<td>5</td>
<td>17.86</td>
</tr>
<tr>
<td>Explanatory Comparative – 4</td>
<td>10</td>
<td>35.71</td>
</tr>
</tbody>
</table>
Fifteen teachers operated at the Comparative Distributional level or higher, which meant that they adequately answered this portion of the task by providing reasonable ranges with appropriate variation for each sample size; ten of these teachers explained the effect of sample size on the variation of sample proportions, which indicated they were functioning at the Explanatory Comparative level. Six teachers functioned at the Pre-Comparative Distributional level, which indicates that although sample size influenced their ranges of acceptable values, they often added in extra “wiggle room” that encompassed an inappropriate amount of variation, presumably to capture all possible values that might occur. Three teachers were coded as Distributional because although their conclusions supported a relationship between variation and sample size, they showed a lack of appreciation for expectation (0.25), and gave ranges where the expected proportion of 0.25 would be considered unusual. A fourth teacher was coded as Distributional because s/he gave the same proportional ranges for both sample sizes, which showed a lack of awareness for how sample size affects the variation of proportions. Three teachers were coded at the lowest level, Inconsistent, because they did not have an awareness of appropriate variation by either providing a point estimate, or by picking arbitrary values that would contain inappropriate variation.

Engaging in Repeated Sampling

To solve this task, each teacher had to decide how many repeated samples to take in order to form their conclusions. Most teachers took the same number of samples for Friday and Saturday each, so if a teacher took 6 samples (for each day), then s/he would be placed in
the 5 to 9 category. All of the teachers who used *Probability Explorer* (with the exception of one that was unclear) simulated between 1 and 20 samples for Friday and Saturday each. The specific breakdown for each category is shown in the bar chart below.

![Bar chart showing the number of repeated samples taken for each day](image)

**Figure 14: Number of Repeated Samples Teachers Took Using *Probability Explorer***

Nine teachers were placed in the “5 to 9” category, but five of those teachers indicated they took “several” samples. Also, one teacher did not indicate how many samples were taken, so this response was not included in the chart. Regrouping based on the coding scheme for high, medium, and low samplers, the majority were medium samplers (61%), while 35% were low samplers. A likely reason that teachers did not take many samples while using *Probability Explorer* is because it is time-consuming to do so. While *Probability Explorer* has the advantage of being user-friendly for running simulations, it is not possible to take a sample instantaneously or to take more than one at a time. One teacher wrote “Ideally I would run
closer to 100 trials representing each day to get a more accurate result, but for time sake I am only running five trials of each.” Teachers using *Probability Explorer* may have had a desire to take more samples, but perhaps did not have the time or energy to do so.

*Use of Probability Explorer to Formulate Conclusions*

Another trend that emerged while analyzing the documents was the types of conclusions that were formed by teachers who used *Probability Explorer*. Thirty teachers included data tables in their screenshots, and these tables appeared to be used most often to help justify their conclusions as supporting evidence. The following frequency table shows the number of teachers who were low and medium samplers, and the types of conclusions they formed based on the samples they took. Recall that low samplers simulated zero to four samples, while medium samplers simulated five to twenty samples. A discussion about the most common conclusion types will follow. If a specific type of conclusion was not found in the *Probability Explorer* documents, then the conclusion was left out of the table.
Table 8: Frequency Table for Types of Samplers vs. Types of Conclusions for Probability Explorer

<table>
<thead>
<tr>
<th>Type of Sampler</th>
<th>The Types of Conclusions Formed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency Percent</td>
<td>Extremes for Any Outcome</td>
</tr>
<tr>
<td>Low</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>7.41</td>
</tr>
<tr>
<td>Medium</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>14.81</td>
</tr>
<tr>
<td>Unclear</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3.70</td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>25.93</td>
</tr>
</tbody>
</table>

Frequency Missing = 4

“Unsupported by Data” was the most common conclusion type for low samplers. This is not surprising because low samplers have taken between zero and four samples for each day, so they have less data to observe. Because of this, many of their conclusions were not supported by the data they simulated. “Extremes for Specified Outcome” was the most common conclusion formed for “Medium” samplers, with ten teachers choosing the most extreme numbers for the hat prize to build ranges of acceptable values. Upon closer examination, eight of these teachers took between 10 and 20 samples to observe the extreme values for hats. These teachers may have desired to take more samples because they were...
only focusing on the hat prize instead of considering all prizes. “High” samplers were not included in this table because no Probability Explorer user simulated more than 20 samples.

Another conclusion type worth mentioning is “Extremes for Any Outcome.” Although “Extremes for Any Outcome” did not win the majority for either “Low” or “Medium” samplers, it is worth noting that the other two technology cases did not have documents that used the data from all of the prizes when building their conclusions; this conclusion only appeared in the Probability Explorer documents. For those responses who considered the extreme values for any prize, it appeared that they were using the data table to consider the data for all of the prizes. These teachers noted that each of the prizes were equiprobable, so you could consider any of the prizes to represent the hat. In effect, they were considering four different values for one sample, so they may have not seen a reason to take more samples.

Microsoft Excel

Twenty-two teachers chose Excel to solve this task which made Excel the second most popular technology choice for this task. The documents ranged from 1 to 15 pages in length, with most documents ranging from 2 to 4 pages in length. Eleven teachers included one or fewer screenshots in their documents, often showing the results of a simulation with a bar chart or pie chart. Thirteen teachers showed spreadsheets of a simulation in their screenshots, seven showed pie charts, and five showed bar charts.

Reasons for Choosing Microsoft Excel

When examining the reasons for choosing Excel, there was less agreement among teachers for top reasons when compared to Probability Explorer. Eight of the twenty-two
teachers who chose Excel did not justify their choice in technology. For those who did offer reasons for choosing this technology, the top reasons are listed in the table below.

Table 9: Top Justifications for Choosing Excel

<table>
<thead>
<tr>
<th>Justification</th>
<th>Frequency</th>
<th>Percentage (out of 22)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple/Easy</td>
<td>8</td>
<td>36%</td>
</tr>
<tr>
<td>Gain Practice</td>
<td>6</td>
<td>27%</td>
</tr>
<tr>
<td>Accessible</td>
<td>5</td>
<td>23%</td>
</tr>
<tr>
<td>Familiarity</td>
<td>4</td>
<td>18%</td>
</tr>
<tr>
<td>Available Representations</td>
<td>3</td>
<td>14%</td>
</tr>
</tbody>
</table>

The most popular justification for choosing Excel is that it is simple and easy to use. Some teachers only stated that it was simple, while others specifically said it was easy to generate random numbers (three teachers) and easy to re-randomize data (two teachers). Others stated that the available functions were intuitive, and the functions embedded in Excel prompt the basic commands needed if the user does not remember completely. Six teachers chose to use Excel to gain practice with the program, seemingly so they would be more equipped to use Excel in their future classrooms. In contrast, four teachers chose Excel because of their familiarity in working with the program. Five teachers stated they wanted to use it because of its accessibility, stating that teachers would be more likely to find this
program in a school setting as opposed to *Probability Explorer* or *Fathom*. Finally, three teachers mentioned that the available representations, such as the pie graph and bar graph, were an advantage to using the program.

*Initial Expectations for the Proportion of Hats*

Teachers were coded as Inconsistent (0), Pre-Consistent (1), Consistent (2), or Distributional (3) based on their responses for part (a) of the task that asked them to list reasonable expectations for the proportion of hats given away for a weekend. One teacher who used *Excel* did not answer this portion of the task, so s/he is not included in the table below.

<table>
<thead>
<tr>
<th>Part (a) Reasoning Levels for Expectation of Proportion for Excel Users</th>
<th>Level</th>
<th>Frequency</th>
<th>Percent (out of 21)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inconsistent – 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pre-Consistent – 1</td>
<td>5</td>
<td>23.81</td>
<td></td>
</tr>
<tr>
<td>Consistent – 2</td>
<td>2</td>
<td>9.52</td>
<td></td>
</tr>
<tr>
<td>Distributional – 3</td>
<td>14</td>
<td>66.67</td>
<td></td>
</tr>
</tbody>
</table>

Fourteen teachers were able to produce a reasonable range of values based on a specific sample size or by reasoning hypothetically, so they were coded at the highest level of Distributional (3). Four of them reasoned hypothetically about different sample sizes, while the remaining ten teachers gave either one range for a specific sample size, or two or more
ranges for various sample sizes. Two teachers were coded one level below at Consistent (2). One teacher was coded at this level because s/he provided too much variation in the range of acceptable values for his/her specified sample size, while the other teacher provided a reasonable range but did not specify an associated sample size. Five teachers functioned at the Pre-Consistent (1) level; four operated at this level because they stated that the proportion of hats would be “about 1/4,” while the other teacher gave one proportion based on a simulated sample. No teachers were coded at the Inconsistent level for Excel users.

*Reasoning About Variation with Respect to Sample Size*

Excel users were also coded based on their responses for part (b) of the task which required them to consider unusual values of hats given away for two different sample sizes. Each response was coded as Inconsistent (0), Distributional (1), Pre-Comparative Distributional (2), Comparative Distributional (3), or Explanatory Comparative (4), as described in the previous chapter. Two teachers did not answer this portion of the task, so they are excluded from the table below.
Table 11: The Reasoning Levels for Excel Users in Part (b)

<table>
<thead>
<tr>
<th>Level</th>
<th>Frequency</th>
<th>Percent (out of 20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inconsistent – 0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Distributional – 1</td>
<td>5</td>
<td>25.00</td>
</tr>
<tr>
<td>Pre-Comparative – 2</td>
<td>2</td>
<td>10.00</td>
</tr>
<tr>
<td>Comparative – 3</td>
<td>6</td>
<td>30.00</td>
</tr>
<tr>
<td>Explanatory Comparative – 4</td>
<td>7</td>
<td>35.00</td>
</tr>
</tbody>
</table>

Thirteen teachers were functioning at the Comparative Distributional (3) level or higher, which means that they gave appropriate ranges of values for both Friday and Saturday that reflected suitable variations of the expectation for the different sample sizes. Seven of these teachers were at the highest level, Explanatory Comparative, which meant that they provided reasonable ranges and made observations about how the sample size affected the variation of the sample proportions. In contrast, two teachers functioned at the Pre-Comparative Distributional (2) level. Both of these teachers provided fairly reasonable responses for what would be considered unusual, but they did not show or explain what data was used to construct those ranges. Five teachers functioned at the Distributional level (1); three of these teachers had the same proportional ranges for both sample sizes, while the other two teachers had more variation of expected proportions for the larger sample size which was not supported by the empirical evidence.
**Engaging in Repeated Sampling**

Two teachers used *Excel* to simulate part (a) of the task, but then based their Friday and Saturday responses on their original expectations of the proportion. Because they did not take samples for Friday and Saturday (even though they took samples for the initial expectations), they were coded as taking zero samples. Thus, *Excel* has two responses that did not take any samples for Friday or Saturday. The bar chart for the other repeated sampling categories is shown below. Three teachers did not indicate how many samples they took (or it was not clear from their responses), so they were excluded from the bar chart.

![Number of Repeated Samples Taken for Each Day](chart)

**Figure 15: Number of Repeated Samples Teachers Took Using *Excel***

Teachers who used *Excel* seemed to use it in dichotomous ways. Six teachers took one or fewer samples for Friday and Saturday each. In contrast, twelve teachers took ten or
more samples for each day. The number of samples taken in Excel appears to be more a matter of personal preference rather than a constraint or advantage inherent in the technology. Many teachers wrote about using the F9 key to simulate new samples quickly, but this feature does not record each simulation. One teacher who recognized the importance of taking many samples wrote “I’m not sure if it was possible to have Excel automatically log all of my trials rather than manually typing them in one by one…” Thus, while it is simple to take multiple samples in Excel, it does not appear to be easy to automatically make comparisons between the simulated samples.

*Use of Excel to Formulate Conclusions*

Of the three technology cases, Excel had the most diverse group in terms of sampling type. Again, certain conclusions were more common depending on the type of sampler the teacher was coded as, and some of the more common conclusions were similar to those discussed in the *Probability Explorer* case. Recall that low samplers simulated between 0 and 4 samples, medium samplers simulated between 5 and 20, and high samplers simulated 21 or more samples. The following frequency table shows the number of teachers who were low, medium, and high samplers, and the types of conclusions they formed based on the number of samples they simulated. Three teachers were considered “Unclear” in terms of what type of sampler they were, but they were included in the table to show the strategies they used.
Table 12: Frequency Table for Types of Samplers vs. Types of Conclusions for Excel

<table>
<thead>
<tr>
<th>Type of Sampler</th>
<th>The Types of Conclusions Formed</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Excluding Extreme Values</td>
<td>Formal Procedure</td>
<td>Extremes for Specified Outcome</td>
</tr>
<tr>
<td>Low</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Medium</td>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>9.52</td>
<td>19.05</td>
</tr>
<tr>
<td>High</td>
<td>4</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>19.05</td>
<td>0.00</td>
<td>9.52</td>
</tr>
<tr>
<td>Unclear</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>9.52</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>19.05</td>
<td>9.52</td>
<td>38.10</td>
</tr>
</tbody>
</table>

Frequency Missing = 1

Five of the teachers who were considered low samplers formed conclusions that were “Unsupported by Data.” This was very similar to low samplers in Probability Explorer, suggesting that this may be a trend specific to type of sampler instead of type of technology. The most common conclusion type for medium samplers in Excel was “Extreme for Specified Outcome,” which considered only the extreme values of the hat prize. While this conclusion type also occurred in Probability Explorer, in Excel it typically occurred because the teachers utilized the “countif” command to only count the hat prizes. By doing this, the problem setup became like a success/failure outcome, with success denoting “hat” and failure
denoting all other prizes. Thus, instead of considering equiprobable results, the majority of teachers using Excel focused only on the hat prizes. For four out of six of the high samplers who used Excel, the conclusion type used was “Excluding Extreme Values.” The teachers who were “High” samplers observed more data, and noted which results were typical and which results came up less often. Because they realized that some extreme results were less likely to occur, they decided to exclude these values from their ranges of acceptable values.

**Fathom**

Four teachers chose Fathom to solve this task, and their documents were the most consistent in terms of length and detail. Three of the documents were five pages in length, while the fourth document was seven pages. Each teacher showed at least one screenshot of a dot plot, while three teachers showed histograms and two teachers showed box plots. Each teacher who used Fathom detailed how they used the program to solve the problem, and they used screenshots as supporting evidence.

**Reasons for Choosing Fathom**

The justifications for choosing Fathom are shown in the table below.
Table 13: Top Justifications for Choosing Fathom

<table>
<thead>
<tr>
<th>Justification</th>
<th>Frequency</th>
<th>Percentage (out of 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graphical Representations</td>
<td>4</td>
<td>100%</td>
</tr>
<tr>
<td>Calculate Statistics</td>
<td>3</td>
<td>75%</td>
</tr>
<tr>
<td>Empirical Sampling Distribution</td>
<td>2</td>
<td>50%</td>
</tr>
<tr>
<td>Speed</td>
<td>2</td>
<td>50%</td>
</tr>
<tr>
<td>Ease of Use</td>
<td>1</td>
<td>25%</td>
</tr>
</tbody>
</table>

The four teachers who chose Fathom all stated the advantages of the graphical representations available. The ability to display results in an organized manner, and depict the results from multiple samples in dot plots was one of the main appeals of Fathom. The functions available to calculate statistics like the mean and standard deviation, and display these on the graph were also appealing to three of the teachers. Two teachers alluded that the ability to generate empirical sampling distributions was a major reason for choosing to use Fathom over Probability Explorer and Excel. Additionally, two teachers believed that Fathom’s ability to perform calculations and simulate data quickly was another advantage. Finally, one teacher stated that Fathom was easy to use; however this was not a consensus among the other Fathom users.
Initial Expectations for the Proportion of Hats

Each teacher was coded based on their response to part (a) of the task as Inconsistent (0), Pre-Consistent (1), Consistent (2), or Distributional (3). Two of the teachers who used Fathom did not answer this portion of the task, so they are excluded from the table below.

Table 14: The Reasoning Levels for Fathom Users in Part (a)

<table>
<thead>
<tr>
<th>Level</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inconsistent – 0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pre-Consistent – 1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Consistent – 2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Distributional – 3</td>
<td>2</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Both of the teachers who answered part (a) of the task operated at the highest level, Distributional, for the initial expectations of reasonable proportions. One teacher took 200 samples of size 8000, and built a reasonable range of values based on that data. This teacher also stated how the range of values might vary based on sample size. The other teacher made a range of reasonable expectations based on 500 samples of size 350. Both of these responses were coded at the highest level because they provided reasonable ranges based on simulated data.
Reasoning About Variation with Respect to Sample Size

For part (b) of the task, teachers were coded as Inconsistent (0), Distributional (1), Pre-Comparative Distributional (2), Comparative Distributional (3), or Explanatory Comparative (4), based on the ranges they provided for the two different sample sizes. The levels for the teachers who used Fathom are shown in the table below.

Table 15: The Reasoning Levels for Fathom Users in Part (b)

<table>
<thead>
<tr>
<th>Level</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inconsistent – 0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Distributional – 1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pre-Comparative – 2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Comparative – 3</td>
<td>2</td>
<td>50.00</td>
</tr>
<tr>
<td>Explanatory Comparative – 4</td>
<td>2</td>
<td>50.00</td>
</tr>
</tbody>
</table>

All four teachers provided reasonable ranges with appropriate variation for the two different sample sizes. Two teachers explained how the variation of sample proportions differed based on the sample size, so they were coded as Explanatory Comparative (4). Each document was very detailed about what data was used to form conclusions which made the coding easier.
Engaging in Repeated Sampling and Forming Conclusions

In contrast to the teachers who used Probability Explorer and Excel, all of the teachers who used Fathom simulated a high number of samples. All four teachers took more than 51 samples, so they would be considered “High” samplers. Because Fathom has the ability to quickly collect and store sample proportions or counts, teachers were able to simulate more samples simultaneously, and keep track of the simulated data without manually recording it. By storing these values and generating dot plots, teachers were able to consider the empirical sampling distributions of the proportions or counts. Since they had more sample proportions to consider, three of the teachers who used Fathom decided to exclude the extreme values to create their ranges. For example, the dot plot below shows the number of hats given away on Friday for 100 simulated samples. This particular teacher stated that if more than 30 or fewer than 18 hats were given away, then this should surprise the manager. However, the dot plot indicates that seven samples yielded values less than 18, and three samples produced values higher than 30, which shows that the teacher decided to exclude the most extreme values to create the range.
One teacher considered the extreme values of hats to build the range of acceptable values. Each teacher included screenshots that showed dot plots of empirical sampling distributions, which made it easier to determine the types of conclusions they formed. They also focused only on the hat prize, evidenced by the way they collected the measures for that prize. In that sense, they approached it in a similar way to many Excel users because they did not necessarily consider the data for other prizes.

Comparing the Cases

The popularity of the available technologies vary greatly when you consider that thirty-one teachers chose to use Probability Explorer, while only four teachers chose to use Fathom. The purpose of this section is to consider how teachers solved the task similarly or differently from teachers who chose a different technology, and in what ways the technology
itself may have constrained or aided in those strategies. Since teachers were not randomly assigned to technology choice, no cause and effect relationship should be implied between technology use and the quality of the responses.

*Reasoning Levels in Parts (a) and (b)*

According to the frequency tables, each technology case had a majority of its teachers operating at the highest level of Distributional (3) for the initial expectations of the proportion. *Excel* and *Probability Explorer* were most similar since both technology cases had about 20% of the teachers operating at the Consistent (2) or Pre-Consistent (1) level each. *Fathom* did not have this in common, but only two of the four teachers who used *Fathom* answered part (a) of the task, so such a small sample size is difficult to compare. The final levels of expectations with respect to sample size followed the same general trends. The majority of teachers in each technology case functioned at the highest levels of Comparative Distributional (3) or Explanatory Comparative (4). Again, *Fathom* teachers were not coded at any of the lower levels, but both *Excel* and *Probability Explorer* had about 25% of teachers operating at the lowest levels of Distributional (1) or Inconsistent (0).

*Comparing Repeated Sampling for Technology*

Technology choice tended to offer either constraints or support when it came to repeated sampling. *Probability Explorer* appeared to limit a teacher’s willingness to take a high number of repeated samples since each sample took time to simulate, and results had to be recorded manually. This was demonstrated by the fact that only ten out of thirty-one teachers took between 10 and 20 samples, and no teachers simulated more than 20 samples.
Additionally, some teachers who used *Probability Explorer* expressed a desire to take more samples, but stated that the time it would take to do so led them to choose a smaller number. In contrast, teachers who used *Excel* had more freedom in the number of samples they could simulate in a shorter amount of time. Although six teachers simulated one or fewer samples in *Excel*, another six teachers simulated 21 or more repeated samples. The ability to use the F9 key to re-simulate data rapidly allowed teachers to collect more samples at a much faster rate. However, these simulations still had to be done one at a time, so teachers could not simulate multiple samples simultaneously (unless they had more than one column to represent more than one sample). Similar to *Excel*, *Fathom* also has the ability to simulate samples instantaneously. However, *Fathom* also has a function that allows users to collect measures, so multiple sample statistics can be simulated and recorded simultaneously. This feature appeared to support the teachers’ use of repeated sampling in *Fathom* since all of the teachers simulated 100 or more samples.

*Comparing Conclusion Types for Technology*

Certain conclusion types were more popular for specific technologies. *Excel* and *Probability Explorer* both had a relatively high number of teachers who formed conclusions that were “Unsupported by Data,” or “Extremes for Specified Outcome,” which meant that teachers formed their conclusions by using the most extreme values for the hats prize. The “Unsupported by Data” conclusion type seemed to be more directly related to the type of sampler a teacher was coded as rather than the type of technology s/he chose. In contrast, many teachers who used *Excel* focused on the extreme values of hats because they often only
counted the number of hats given away rather than the other prizes. While teachers also did this in *Probability Explorer*, another conclusion type, “Extremes for *Any* Outcome”, was popular only for this specific technology. Teachers who were coded as “Extremes for *Any* Outcome” formed their conclusions by considering the extreme values for *any* prize rather than only considering the data for hats. This appeared to happen in *Probability Explorer* rather than the other technologies because of the available representation of the data table. The data table allowed the teachers to consider the outcomes of all four prizes, and since many of the teachers recognized these were equiprobable, they realized that any prize could represent the hat. This conclusion type did not occur in other technologies mostly because teachers only collected data for hats, and no other prizes. *Fathom* and *Excel* shared a common conclusion type of “Excluding Extreme Values”, which meant that teachers excluded the most extreme values or proportions of hats. This appeared to be most closely related to the number of samples teachers simulated because a higher number of samples tended to be associated with this conclusion type.

*Choosing a Technology*

*Probability Explorer* was the most popular technology choice in this study seemingly because of its dynamic nature and ease of use. Compared to *Excel* and *Fathom*, it is much easier to visualize a sample being collected and to see how the simulations affect the available representations (pie graph, bar graph, data table) as they occur. Designing an experiment is also more intuitive, as well as creating the associated graphs. *Excel*’s main advantage was its ease of use because it was easy to simulate results, and if the user
remembered the name of a formula, then it could prompt the user on how to fill in the rest. Additionally, Excel has the benefit of being more readily available in a school setting, so teachers may have wanted to gain more practice since they might actually have that technology in their classroom. Fathom’s clear advantage is its ability to take many samples and store that information for the user to analyze. However, learning how to do this is not as simple as learning how to use Probability Explorer or Excel. One teacher expressed this frustration when s/he wrote “The only hindrance I encountered was trying to generate the random numbers needed. In Probability Explorer, Excel, and the calculator the steps to generating a simulation are less complicated.” Another teacher echoed this sentiment by stating “I am not as comfortable with Fathom. Just to create this single random variable, I had to go online to look up some steps…” Based on the fact that Fathom was the smallest collection of responses and the teachers who used it expressed frustration, it is my supposition that the teachers who chose it were purposeful in their decision to use it in order to take more samples to solve the problem.
The objective of this chapter is to consider the relationship between the type of sampler and the reasoning levels, as well as to compare the reasoning levels in part (a) and part (b) to one another. Throughout the analysis of the cases, it became evident that the reasoning level a teacher was coded as was probably less influenced by the technology s/he chose, and more influenced by the type of sampler s/he was. Because of this, I chose to consider how the type of sampler was associated with the levels for variation with respect to sample size, as well as observing the aggregate data for the types of conclusions formed. Additionally, it appeared that many teachers focused more sophisticated strategies on one portion of the task, part (a) or part (b), and as a result had different reasoning levels for the different parts of the task. Thus, it was necessary to do a more close analysis to observe the relationship between the reasoning levels for the different parts of the task.

Type of Sampler vs. Variation with Respect to Sample Size

The levels for types of sampler are based on the number of repeated samples they took for Friday and Saturday each. Since I did not document the number of samples they took for the initial expectation of the proportion (which was often different from the number of samples they decided to take for Friday and Saturday), I am only going to consider the results for the types of sampler vs. the final levels of variation with respect to expectation and sample size. A frequency table for the types of sampler vs. the final levels of variation with respect to sample size is shown below.
Table 16: A Frequency Table for Type of Sampler by Levels for Comparing Variations

<table>
<thead>
<tr>
<th>Type of Sampler</th>
<th>Frequency</th>
<th>Row Percent</th>
<th>Inconsistent 0</th>
<th>Distributional 1</th>
<th>Pre-Comparative 2</th>
<th>Comparative 3</th>
<th>Explanatory Comparative 4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>2</td>
<td>14.29</td>
<td>2</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>Medium</td>
<td>1</td>
<td>4.17</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>7</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4.17</td>
<td>4.17</td>
<td>20.83</td>
<td>29.17</td>
<td>41.67</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>50.00</td>
<td>50.00</td>
<td></td>
</tr>
<tr>
<td>Unclear</td>
<td>0</td>
<td>0.00</td>
<td>1</td>
<td>25.00</td>
<td>0.00</td>
<td>0</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.00</td>
<td>25.00</td>
<td>0.00</td>
<td>0.00</td>
<td>75.00</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3</td>
<td>9</td>
<td>8</td>
<td>13</td>
<td>19</td>
<td></td>
<td></td>
<td>52</td>
</tr>
</tbody>
</table>

Frequency Missing = 5

From the table, it is observed that nine of the fourteen low samplers (64% of the low samplers) are operating at Distributional (1) or lower. Distributional (1) indicates that teachers only attended to two of the three aspects of the problem (variation, sample size, and expectation). This typically manifested itself in the same proportional ranges for both Friday and Saturday, indicating a lack of appreciation for how sample size affects variation. Given that the teachers who were “Low” samplers took a maximum of four samples, they likely did not collect enough data to notice how the sample proportions varied for each sample size. Teachers operating at the Inconsistent level indicated a lack of awareness about variation in
general with the belief that anything can happen, or it should be exactly 25% for each sample size. Only two of the low samplers (14%) were functioning at the Comparative Distributional (3) level or higher. Upon further investigation, it was discovered that the two low samplers at the higher levels took between 2 and 4 samples instead of 0 or 1 sample.

The medium samplers tended to operate at higher levels than the low samplers, with seventeen of twenty-four teachers (71% of the medium samplers) operating at the Comparative Distributional (3) level or higher. Teachers at this level provided reasonable ranges for both Friday and Saturday based on the data they simulated. The teachers who went above and beyond this by making correct statements about how the sample size affects the variability of sample proportions were coded as Explanatory Comparative. However, five of the medium samplers (21% of medium samplers), were functioning at the Pre-Comparative Distributional (2) level. Teachers working at this level either gave ranges that had a little too much variation around the expectation which encompassed some values that should be considered unusual, or they gave reasonable range(s) without data to support how they found them.

The high samplers were the most consistent group, with all teachers functioning at the Comparative Distributional (3) level or higher. Since the task did not specifically ask how the sample proportions varied based on sample size, those teachers operating at the Comparative Distributional level successfully answered the problem, and those operating at the Explanatory Comparative level exceeded the requirements of the task. When we consider the aggregate data, of the fifty-two teachers who were coded by the levels, thirty-two of them
(62%) answered the second portion of the task satisfactorily. Eight teachers (15%) operated at the Pre-Comparative Distributional Level, so with more training about appropriate variation and using data to support conclusions, they are on the cusp of successfully solving an informal inference task such as this. However, twelve teachers (23%), operated at the Distributional or Inconsistent level for the second portion of the task. This is concerning because it suggests a lack of understanding of how the topics of variation, sample size, and expectation are interwoven, and knowledge of these topics is vital in order to support future students’ informal inferential reasoning.

*Type of Sampler vs. Type of Conclusion*

Another trend that emerged during the analysis is that the type of sampler (low, medium, high) was often associated with certain types of conclusions. The most common conclusion types were “Unsupported by Data,” “Extremes for Any Outcome,” “Extremes for Specified Outcome,” and “Excluding Extreme Values.” Recall that “Unsupported by Data” means that the conclusions were not supported by the data mentioned or revealed through screenshots. “Extremes for Any Outcome” refers to the teachers who formed their conclusions based on the extreme values of any prize, either by explicitly stating that they were doing so, or as inferred through screenshots. “Extremes for Specified Outcome” denotes the conclusions that were formed based on the extreme values of only the hat prize, without considering any other prize results. Finally, “Excluding Extreme Values” indicates that the teacher excluded the most extreme values typically because they considered those
values to be “unusual” since they did not occur often. A frequency table of type of sampler vs. type of conclusion is shown below.

Table 17: A Frequency Table for Type of Sampler by Type of Conclusion

<table>
<thead>
<tr>
<th>Type of Sampler</th>
<th>Frequency Row Percent</th>
<th>The Types of Conclusions Formed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Extremes for Any Outcome</td>
<td>Multiple Strategies</td>
</tr>
<tr>
<td>Low</td>
<td>2</td>
<td>13.33</td>
</tr>
<tr>
<td>Medium</td>
<td>4</td>
<td>17.39</td>
</tr>
<tr>
<td>High</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>Unclear</td>
<td>1</td>
<td>25.00</td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>2</td>
</tr>
</tbody>
</table>

From the table it is apparent that most of the low samplers (73%) did not support their conclusions with data. Given that justifying conclusions with data is an important facet of informal inferential reasoning, it is alarming that twelve teachers (total) did not feel the need to use data as evidence to support their conclusions. In contrast, the most popular conclusion
type for medium samplers (61%) was to build ranges of acceptable values based on the extreme values of hats that were simulated. Of the high samplers, 70% chose to exclude the extreme values to build acceptable ranges. Excluding extreme values begins to lay the foundation for confidence intervals and hypothesis tests, so in a sense, this might be considered one of the more sophisticated ways of forming conclusions. The high samplers may have been more comfortable with excluding extreme values because they observed more sample proportions, and if they observed an extreme value that was farther away than the others, then they could deem it “unusual” with more confidence. In comparison, if a medium sampler only took ten samples and noticed extreme values, s/he does not have as many values to compare it to, so s/he may not realize how extreme it really is in relation to the other sample proportions. Furthermore, it may be a matter of personal preference to consider all of the simulated values to be reasonable since they were actually observed, and use the extreme values as buffers between what is reasonable and what is unusual.

*Comparing Levels of Reasoning for Initial Proportions and Comparing Variations*

By reading through the responses several times, it became apparent that while some teachers were thorough in answering both (ii) of part a, and part b, many other teachers focused more energy on one of the parts. Because of this, a teacher’s level of reasoning on making an initial prediction of an expected proportion and their reasoning when comparing variations from different sample sizes do not match up perfectly. Some teachers scored better in making their reasonable expectations for part (a), while others scored better on comparing variations from different sample sizes for part (b), which seems counter-intuitive. The
frequency table below shows the breakdown of the reasoning levels for parts (a) and (b) of the task.

Table 18: Comparing Levels of Reasoning in Part (a) and Part (b)

<table>
<thead>
<tr>
<th>Part (a) Reasoning Levels for Expectations</th>
<th>Part (b) Reasoning Levels for Comparing Variations of Two Different Sample Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency Percent</td>
<td>Inconsistent</td>
</tr>
<tr>
<td>Inconsistent</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.75</td>
</tr>
<tr>
<td>Pre-Consistent</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.72</td>
</tr>
<tr>
<td>Consistent</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>Distributional</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.75</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>Total</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>5.26</td>
</tr>
</tbody>
</table>

The frequency table is color-coded to show which teachers improved from part (a) to part (b) of the task, which stayed at the same type of level, and which did not do as well in part (b) compared to part (a) of the task. The blue color indicates improvement, the peach color indicates no change, and the red color indicates a downgrade in levels. The levels for part (a) and part (b) do not match up perfectly, so some of the levels for part (a) correspond to two different levels for part (b) of the task. We would expect that teachers operating at the Inconsistent level in part (a) would continue to stay at the Inconsistent level for the second
part of the task because these levels include most of the same criteria. The Pre-Consistent level does not match up as well to the final levels. A teacher operating at the Pre-Consistent level could conceivably be at the Inconsistent level for part (b) of the task if s/he only provided a point estimate for the initial expectations of the proportion. However, if s/he stated that it would be “about 25%” without a corresponding sample size, then s/he may be at the Distributional level for part (b) of the task by attending to only two of the three aspects of the problem (expectation, variation, or sample size). “About” indicates a general awareness of variation, and “25%” shows an awareness of expectation, but there is no clear attention to the role of sample size. Teachers operating at the Consistent level in part (a) could be expected to operate at either Distributional or Pre-Comparative Distributional for part (b). If teachers were initially coded as Consistent because they provided reasonable ranges of expectations without indicating the sample size or data used to construct the ranges, then it is possible that they may be at the Distributional range by providing the same proportion ranges for Friday and Saturday (lack of attention to the role of sample size). However, if teachers were coded as Consistent because they initially gave a range of proportions that encompassed too much variation, then they may be Pre-Comparative Distributional because teachers at this level tended to give too much extra “wiggle room.” Finally, if teachers were operating at the Distributional level in part (a), then it would be expected that they would operate at either Comparative Distributional or Explanatory Comparative. They should continue to operate at the highest levels because if they were coded as Distributional for thought experiments based on sample size, then they obviously have an awareness of the role of sample size. If teachers
were coded as Distributional because they gave appropriate ranges around 25% based on specific samples, then they should be able to do the same in part (b) of the task for the two different sample sizes.

Five teachers improved from a Pre-Consistent to a Pre-Comparative or Explanatory Comparative level. Of the teachers who improved to the Pre-Comparative level, they did so largely because they gave somewhat reasonable range(s), but did not have data to support what they did. For the teachers who jumped from the Pre-Consistent level to the Explanatory Comparative level, both were coded at the Pre-Consistent level in part (a) because they stated that the expected proportion should be “roughly ¼”. These teachers did not give an associated sample size or simulated data, so they were coded as Pre-Consistent. Although they did not place much emphasis on the first part of the task, both teachers went above the requirements for part (b) by making observations about how the variability of sample proportions was affected by the sample size use. Similarly, four other teachers did not answer the first part of the task (their initial levels were coded as missing), but they were also coded as Comparative or Explanatory Comparative for the second part of the task. These teachers are examples of how their initial levels may not actually be indicative of what they know with respect to sample size, variation, and expectation.

Thirty-one teachers stayed at their expected levels from parts (a) and (b) of the task. This makes up about 54% of the teachers in this study, which is somewhat surprising considering that one might predict the teachers would generally stay around the same level. Seven teachers (12%) went down in levels from part (a) to part (b), while another three
teachers did not answer the second part of the task at all, making it impossible to rate them. Most concerning, perhaps, is the fact that four teachers had the highest initial level of Distributional (3) and downgraded to the Inconsistent (0) or Distributional (1) level for part (b) of the task. This is worrisome because the teachers initially gave an appropriate range for the expectation of the proportion based on sample(s) or sophisticated thought processes, but for part (b) of the task they did not consider how the different samples sizes for Friday and Saturday would affect the proportions in the ranges of acceptable values. One teacher in particular simulated 100 samples of size 5000 to find reasonable values for the expectation of the proportion, but then there was no indication s/he took any samples for Friday or Saturday. It appeared that this teacher used the initial range of proportions to construct the same proportional ranges for Friday and Saturday, only using the sample size to find the number of hats. It is interesting that this teacher would use such a sophisticated approach to answer the first part of the question without considering what differences might occur when different sample sizes are used.

The purpose of comparing reasoning levels was to observe whether teachers were consistent in both parts of the task. About half of the teachers operated consistently, but the other half fluctuated depending on which part of the task they were answering. Part of this appears to be based on which part of the task they found more important – finding reasonable expectations for the proportion of hats given away, or determining unusual values for two different sample sizes. Some teachers had a solid understanding of the relationship between the three aspects, but did not consider the first part of the task very deeply. In contrast, other
teachers had refined strategies for determining an appropriate range of acceptable values for the expectation of proportion, but then became too confident with this initial range, and used it for the different sample sizes.
The purpose of this study is to consider teachers’ approaches to an informal inference problem. Informal statistical inference is a burgeoning topic that teachers will be expected to teach, with the goal of teaching students to make well-reasoned inferences using data as evidence (Makar & Rubin, 2009; Zieffler, et al., 2008). In order to collect this data and explore these statistical topics, both teachers and students need to be able to use technology to run simulations (Drier, 2001; Rossman & Chance, 1999; Ben-Zvi, 2000). By considering the interplay between technology choice and the types of conclusions formed on informal inference tasks, we can consider which technologies might support certain statistical concepts. Additionally, by observing the processes which a teacher takes to solve a task, we can make teacher education programs better to support teachers’ understandings of the relationship between expectation, variation, and sample size. To address these issues, this chapter seeks to summarize the main findings in this study, and will offer suggestions for future research, as well as limitations inherent in the study. This chapter is divided into four sections: summaries and conclusions, limitations, implications, and recommendations for future research.

Summaries and Conclusions

This study analyzed the documents of fifty-seven pre-service and in-service teachers enrolled in classes that focused either on using technology to teach middle or high school mathematics, or teaching statistics in elementary or middle school classes. Teachers chose
either *Probability Explorer, Excel,* or *Fathom* to solve an informal inference task, and each document described their use of the technology in solving the problem. Because technology played a central role in how teachers solved the task, each technology was considered to be a separate case, with the goal of drawing comparisons between the choice of technology and how the task was solved. A modified version of the framework by Watson et al. (2007) was used to code each response based on their initial answers of reasonable expectations of the proportion, as well as a separate code for their ranges of values for Friday and Saturday. Because their levels were often related to the number of samples the teachers took, an open coding scheme was used to code teachers as low samplers (0 to 4 samples), medium samplers (5 to 20), and high samplers (21 or more). Additionally, an open coding scheme was used to code the ways in which teachers formed their conclusions. The framework and coding schemes were used to answer the following research questions:

1. How do teachers use different technology tools to conduct simulations and make informal inferences?
2. How do teachers reason about variation with respect to expectation and sample size, and how do they use data to develop reasonable predictions?

*Research Question 1*

Each technology case had responses where teachers simulated repeated samples of the same size, but the number of repeated samples simulated was not the same across the technologies. Teachers who used *Probability Explorer* tended to take fewer samples than teachers who used other technologies. Teachers in this case took a maximum of 20 repeated
samples for each sample size. This probably stemmed from a constraint in the technology; *Probability Explorer* does not have a feature that allows the user to simulate new samples instantaneously, and it does not automatically record the results that have already occurred. In contrast to *Probability Explorer*, *Excel* users varied extensively in the number of samples they chose to take for each day. Six teachers chose to take one or fewer samples, while twelve teachers chose to simulate ten or more samples for each day. For those teachers who did perform multiple simulations, they mentioned using the F9 key which re-randomizes the simulated results instantly in *Excel*. This feature allowed teachers to simulate multiple samples rapidly; however, it did not record these results, so teachers had to take mental notes or manually record the outcomes. Teachers who used *Fathom* were most consistent in terms of repeated sampling. After simulating the data, each teacher collected the measures from 100 or more repeated samples, and displayed the results in a dot plot. The ability to collect many samples at once and store the simulated results was a clear advantage in *Fathom*.

The formation of conclusions had both similarities and differences across the cases of technologies. Teachers who used *Probability Explorer* formulated their conclusions by three main strategies: using the extreme values for *any* prize, using the extreme values for *only* hats, or by choosing values that were unsupported by data. The first option of choosing the extreme values for *any* prize was observed only in the *Probability Explorer* documents; users in *Excel* mentioned that other prizes could be considered, but they did not use the data from other prizes to actually build their ranges. One of the visual representations in *Probability Explorer*, the data table, seemed to be most often linked with this type of conclusion.
formation because it showed the numbers and percentages for each prize. While seven teachers considered the extreme values for any prize, the most common conclusion type for teachers using *Probability Explorer* was to consider the data only for hats. This meant that the teacher simulated two or more samples, determined the most extreme values for the hats, and then based their ranges of acceptable values upon those numbers. In effect, these teachers were stating that lower or higher values than the most extreme numbers for hats would be considered “unusual.” This was also the most common conclusion type for teachers who used *Excel*. Considering only hats seemed directly related to how teachers used *Excel* because most teachers only counted the prize if it represented the hat. By considering only the hats, this hints at a binomial distribution in which hats are considered a “success” and all other prizes are considered a “failure.” This type of reasoning would be especially useful if the prizes did not have equal probabilities of occurring. Two teachers chose to take advantage of the more formal statistical functions embedded in *Excel* by using the chi-square test function, as well as the average standard deviation functions. Each teacher still simulated the results for different sample sizes, but they used those simulations in conjunction with the formal statistical measures.

The most common way that the *Fathom* teachers formed conclusions was by excluding the most extreme values. Two teachers stated a range of acceptable values, and then said that observed extreme values were actually outside the given range, suggesting that these teachers felt comfortable excluding values that did not occur often. Another teacher used a formal statistical measure, standard deviation, to determine where most of the data
fell. S/he chose to exclude the values that were not within two standard deviations of the mean. Only one teacher chose to include the most extreme values in his/her range, which may have been a matter of personal preference. However, it is worth noting that this teacher collected 500 samples for each day (as opposed to about 100 for the other teachers), so s/he may have observed a larger quantity of extreme values than the other teachers, potentially resulting in a reluctance to exclude those values.

Each technology choice has advantages and disadvantages inherent in their designs. While *Probability Explorer* is a simulation tool that many teachers described as easy to use, with the ability to depict simulations as they are occurring, it may not have been the best choice for this task. Given that the nature of this task prompts teachers to consider data from multiple samples, the inability to take more than one sample at a time or to re-randomize instantaneously is a major drawback. Also, while the data table allowed teachers to consider data for other prizes, if this task had not had equiprobable results, then this might have actually hindered more than helped. In contrast to *Probability Explorer*, Excel’s ability to re-randomize simulated results makes it a much more desirable tool for repeated sampling.

However, one drawback of *Excel* is that you cannot visualize the simulations as they occur, so this tool may not be the best to introduce the topic of computer simulations. Both *Probability Explorer* and *Excel* were described by several teachers as “easy to use,” but teachers who used *Fathom* did not state that the tool was simple. Two of the teachers using *Fathom* stated that they had difficulty remembering the commands in this particular tool, which could explain why it was not as popular a choice as the other technologies. However,
each of these teachers used *Fathom* to collect many samples, suggesting that they may have chosen this technology because of its power to do repeated sampling. Also, each teacher who used *Fathom* included a dot plot in their screenshots, which was an unavailable feature in the other programs. This graphical representation allows users to consider empirical sampling distributions, which helps build the foundation for many important statistical topics.

When teachers choose to use technology in their classrooms, they have to consider what is available, how much time the technology will require in order to become familiar with it, and which technology will help illuminate the desired concepts most clearly.

*Probability Explorer, Excel, and Fathom* could each be used in a school setting, with some settings more appropriate than others. In spite of the fact that its computing power is not as high as other available technologies, *Probability Explorer* is an excellent tool that helps users design experiments and visualize simulations as they happen. The ease of use makes this tool desirable for teaching basic probability concepts to children in elementary and middle school settings because it does not require much time to learn how to use. This tool could also be used in a high school setting if the task was designed appropriately, and perhaps considered the data from multiple students instead of considering the results from only one user. *Excel* is a desirable simulation tool because it is widely available, and it has more formal statistical measures available. This could be considered an advantage of *Excel* because users can keep an experiment as simplistic as desired, or they can use the available features of the program to solve the problem in other ways. While using these functions may not be the best idea for building informal inferential reasoning, it could help students make the connections between
informal statistical topics and formal statistical topics if they use a program with which they are already familiar. Fathom is probably the most powerful technology tool for illuminating statistical topics, and it would probably be used best in an introductory statistics class. If a teacher’s goal is to lay the foundation for sampling distributions, confidence intervals, and hypothesis tests, then Fathom is the best tool to use. However, given the time it takes to learn the tool, it should be used either extensively in the class, or perhaps as a demonstration tool. If it was updated to be more intuitive, it might be used more often in the classroom.

Research Question 2

The ways in which teachers reasoned about variation with respect to expectation and sample size appeared to be directly related to the number of repeated samples they simulated. Teachers who collected more repeated samples for Friday and Saturday each tended to form better conclusions because they had more data to consider what was typical and unusual. In contrast, teachers who collected fewer repeated samples for Friday and Saturday often had trouble with at least one facet of the problem, usually with sample size or occasionally with expectation.

Teachers who were considered “low” samplers (n = 14), or those who had taken between 0 and 4 samples, typically rated at the Distributional (1) level for part (b) of the task, with a few above and below that level. This is not surprising because in order to score a three or higher (Comparative Distributional or Explanatory Comparative), a teacher must supply ranges of reasonable values that were based on data. Since the low samplers did not collect much data for Friday and Saturday, they typically did not support their conclusions with data.
In general, they also did not recognize that the sample proportions varied more when the sample size was smaller, so they had difficulty recognizing the relationship between variation and sample size. Because of this, seven teachers operated at the Distributional level which meant they attended to only two aspects of the problem instead of considering the relationship between expectation, variation, and sample size altogether.

“Medium” samplers were teachers (n = 24) who took a minimum of 5 and a maximum of 20 samples. These samplers had much better ratings than low samplers, but 7 out of 24 teachers still scored at the Pre-Comparative Distributional (2) level or below. For the seven teachers operating at this level or below, at times they would use procedures that did not make sense, or they would focus on the extreme values of hats and then widen that range to encompass more variation than was reasonable, presumably to capture all types of sample proportions that might come up. For the seventeen teachers operating at the Comparative Distributional (3) level or the Explanatory Comparative (4) level, they typically focused on the extreme values for any prize, or the extreme values just for hats to build the ranges of acceptable values. They used the extreme values to be the cutoff points between what was considered usual and unusual, and this strategy generally helped them recognize that the variation of sample proportions for Saturday would be less than that for Friday because the sample size on Saturday was larger.

Teachers (n = 10) who simulated 21 or more samples were considered “high” samplers. Five of these teachers operated at the Comparative Distributional (3) level, and the other five at Explanatory Comparative (4), the highest possible levels. By taking more
samples, it appears that teachers were able to produce reasonable ranges of values supported by data, and they often noted that higher sample sizes yield less variation in sample proportions. Of the ten teachers who were high samplers, seven formed their conclusions by excluding the most extreme observed values from their ranges of reasonable hats given away. This type of strategy supports the topics of confidence intervals and hypothesis tests because it considers how likely it is that a value will be observed. A potential lurking variable however, is that teachers with a stronger statistical background may realize the importance of taking more samples and excluding the most extreme values, so perhaps their conclusions trended this way both because of the data they simulated and their knowledge of statistics.

While many teachers did not change reasoning levels from part (a) to (b), sixteen teachers did fluctuate in levels. This number is considerable because you would predict that teachers would probably operate at the same level from part (a) to part (b) of the task. It appeared that teachers changed levels mostly because of how they placed importance on the different parts of the task. Some teachers had their levels go up from part (a) to (b) because they wrote more details and justified their conclusions with data for the second part of the task rather than the first part. Most of the teachers who fit this description stated that reasonable expectations would be about \( \frac{1}{4} \) or 25\%, but they did not state an associated range of proportions. However, in the second part of the task, they found specific ranges for both Friday and Saturday, and based these ranges on their respective sample sizes. This strategy allowed them to go up in levels. In contrast, for teachers who went down in levels from part (a) to (b), these teachers often showed the data used to justify their answers for part (a), and
then wrote conclusions for part (b) without explaining how they found them. Often it appeared that they used their range from part (a) to determine the same proportional ranges for both Friday and Saturday in part (b). In doing so, they lacked an awareness of the role of sample size, and this caused them to be rated at a lower level for part (b).

**Limitations**

The first limitation for this study is that teachers chose their own technology, so technology choice was not randomized. Because teachers chose what technology they wanted to use, one group (*Probability Explorer*) had many more responses than another group (*Fathom*). Some teachers chose the technology that they felt most comfortable with, while others chose technologies to learn more about them, or because they would be more readily available in a school setting. This limits the generalizations one can make based on technology choice because although some technologies allow for sophisticated ways of solving the problem, if the teacher did not know how to use the technology, then they probably chose a different one. Thus, it is difficult to predict whether they would make the same or different conclusions using another technology tool, or if they would even use it in an appropriate way if they did not have enough knowledge about the technology itself. This consideration is important because while one technology may be better for teaching informal inference than others, if teachers do not know how to use it effectively, then it will not be useful in the classroom.

Another limitation in this study was the quality of the written documents the teachers submitted. These documents varied widely in terms of both length and depth. While teachers
may have been verbose on some aspects of the task, they were often vague on other aspects, or did not answer all parts of the question. In order to include as many documents as possible, some assumptions had to be made for those who did not explain themselves entirely, either through close examination of screenshots, or by drawing conclusions based on general impressions. Obviously assumptions are not always made correctly, so it is possible that a few of the documents may be coded inaccurately.

The final limitation in this study is that teachers likely had various mathematical and statistical backgrounds, and since this information was not collected, no conclusions related to their background can be made. Based on the responses it appeared that teachers had very different backgrounds in statistics. Some teachers clearly stated that they used formal statistical procedures or measures to help solve the problem, which is not a good indication of how they would teach the problem informally. Furthermore, if teachers knew of the importance of taking more samples because of background statistical knowledge, then that information may have influenced their choice in technology.

**Implications**

Given the results of this study, as well as the documented difficulties teachers have with learning about collections of samples, it is necessary to have both further research on this topic, as well as further training for teachers (Thompson, Liu, Saldanha 2007; Wild et al. 2011). If teachers do not have adequate knowledge of these foundational statistical topics then it is going to exacerbate the misconceptions that students already have while learning them. According to the Common Core State Standards (2010), students will be expected to
“understand statistics as a process for making inferences about population parameters based on a random sample from that population” (p. 43). In order to make inferences with a measure of confidence, however, students will need to have knowledge of how statistics vary from one another (Watson & Kelly, 2004). Thus, teachers need to be adequately trained on these topics in order to successfully teach students.

In addition to content knowledge, teachers also need to have knowledge of available technologies in order to determine which is most appropriate for the task at hand. In this study, *Fathom* was the most powerful computing tool available, but the vast majority of teachers did not choose this tool. Of the few that did, two stated that they had difficulty remembering the appropriate commands, suggesting that further training in different technologies would be useful. However, the time it takes to learn a tool should be considered before implementing that tool in a classroom setting, since you do not want to spend more time teaching about the technology than the statistical topics themselves.

Many of the observations in this study were in agreement with previous literature. While thirty-two of the teachers operated at the highest levels of Comparative Distributional and Explanatory Comparative, twenty teachers were functioning at the Pre-Comparative Distributional level or lower. This indicates that the teachers had difficulty in coordinating the aspects of expectation, variation, and sample size, which is not too surprising considering the low number of students who operated at the Comparative Distributional level in the Watson et al. study (2007). Teachers in this study actually had much higher success rates for the higher levels, but that may be attributed to age, completion of a data analysis unit, and/or
perhaps the ability to run simulations before forming conclusions. Additionally, teachers functioning at the lower levels of reasoning for comparing variation with respect to sample size were often coded at the lower levels because their ranges encompassed too much variation instead of too little, which has been observed in previous studies (Watson & Kelly, 2004). Teachers were also coded at the lower levels if they did not support their conclusions with data, which may be indicative of their beliefs of the importance of justifying conclusions (Leavy, 2010).

While many of the findings are in accordance with prior studies, some of the results appear to be new. For example, while several studies have focused on the effect of drawing larger samples in order to make stronger conclusions, not many studies have considered how teachers use repeated sampling to form appropriate conclusions. By considering the type of sampler a teacher was (low, medium, or high), it became easier to see how they formed their conclusions in terms of using extreme values or by excluding extreme values. Furthermore, by considering the ways in which teachers attended to features of the technology, it became apparent that certain displays can help them observe equiprobable results, while other displays may focus on a success/failure outcome.

This study also modified the framework by Watson et al. (2007) to fit the nature of this task, and recognized that the two parts of the task may not be emphasized equally by teachers, so they may not be coded at the correct level of understanding depending on which part they think is more important. In order to understand what teachers know about variation
and expectation with respect to sample size, it may be useful to ask more direct questions about the proportional ranges for two different sample sizes.

Recommendations

The results lead to several recommendations for future research. The first recommendation is to change the problem set-up slightly by having three prizes instead of four, with five cards total. One card would represent the hat, and two cards each could represent two of the other prizes. This would give the hat a 20% chance of being chosen, while the other two prizes would have a 40% chance each of being chosen. The reason for this recommendation is that in the original problem, every prize had the same chance of being chosen, so several teachers considered the empirical probabilities for all prizes in their analysis. While this is not necessarily incorrect in the original scenario, it would be incorrect in the modified problem. Additionally, teachers may be inclined to take more samples if they are not able to consider the results of other prizes, but this is a hypothesis that would need to be tested.

Other recommendations I would make are to have more direct questions, either in the problem set-up itself, or in follow-up interviews. For example, I would explicitly ask teachers to compare the variability for Friday and Saturday ranges, and I would ask them to state exactly how many samples they took for Friday and Saturday each. The purpose of asking more directed questions would be to have more information from what the teachers did instead of making educated guesses. This would probably take the form of post-interviews so that I could have a chance to skim the responses and then ask clarifying questions if their
responses were not clear.

Another recommendation which may be more difficult to execute would be to randomly assign technologies to the teachers. The purpose of doing this is twofold. First, this would allow for equal (or comparative) group sizes among the technologies. Second, this would showcase both advantages and drawbacks of different technologies depending upon their ease of use and how comfortable the teacher is with using it. The risk of doing this is that teachers may struggle to solve the problem using a technology that is not of their choice (even though they may have a sophisticated method of solving it). In spite of this risk, I think it would be useful to observe how teachers use a technology that they may not be comfortable with because teachers often do not have choices about which technologies are available in their classrooms.

Furthermore, I would recommend doing pre-questionnaires and post-interviews on this task. The pre-questionnaire could ask general questions, such as how many samples the teacher plans on taking (and why they think that number is appropriate), what their predictions are for the Friday and Saturday ranges, and to describe their backgrounds in statistics. It could also ask the teacher which technology they think would be best to use and why, so the researcher can note whether the teacher was randomly assigned to the technology of his/her choice. The post-interview could ask the teachers follow-up questions, like whether they changed their strategy of solving the problem, or whether their initial predictions were correct. Additionally, questions asking the teacher to compare different technologies and how they could be used to solve the problem may help illuminate trends associated with the
technologies. The purpose of doing a pre-questionnaire, the problem itself, and a post-interview would be to triangulate the data to make the findings more thorough.

The teachers in this study completed a data analysis unit as part of their course, and while they were largely successful in this task, there is still room for improvement. To support informal inferential reasoning, teacher education programs should emphasize the relationship between variation, expectation, and sample size, as well as supporting conclusions with data. Teachers should be instructed about the importance of taking multiple samples in order to observe sample-to-sample differences, and they should also be given the opportunity to compare sample-to-sample differences for different sample sizes. Since recognizing the influence of sample size on the variation of sample statistics is not necessarily intuitive, teachers should make predictions about what they think will occur, and then test their predictions in order to confront any misconceptions they may have (delMas, Garfield, & Chance, 1999). Additionally, teachers should be taught how to test conjectures by running simulations, and they should be instructed to use the results of the simulations as supporting evidence for their conclusions. Many of the teachers in this study did not make it explicitly clear how they were using the results of their simulations to form their conclusions, so it may be necessary to emphasize using data as supporting evidence in a teacher education program.

This study utilized two frameworks, each modified from the framework described in the study by Watson et al. (2007) to code the reasoning levels of teachers for two different parts of the task. The frameworks were modified partially based on what was observed
throughout the documents, which may not include all possible responses for this task. The reasoning levels of teachers did not always match up perfectly from one part of the task to the next, and this may be remedied by asking teachers more clarifying questions in a post-interview. Throughout the process of modifying the frameworks, attention was focused on how teachers were reasoning in forming their conclusions, and which types of reasoning were more sophisticated than others. The proposed frameworks attempt to classify each response to determine whether the teacher was operating at a higher or lower level of informal inferential reasoning. By doing so, it is easier to observe where gaps of knowledge may be for teachers. The adapted framework for this study could be used in other tasks of a similar nature that deal with expectation, variation, and sample size. A few levels were added in (such as Pre-Consistent, and Pre-Comparative Distributional) to try to pinpoint the specific levels of reasoning a person should be coded as. The levels of reasoning are generalized so that it could be altered to fit tasks that consider sample means instead of sample proportions, or to tasks that do not have equiprobable results. This type of framework could also be used in a classroom; for example, if students are not recognizing the effect of sample size on the variation of statistics, then it may be necessary to do more direct instruction on this topic. Also, when using these frameworks, it will be beneficial to ask clarifying questions in order to correctly gauge the appropriate reasoning level.

The subjects of mathematics and statistics are often considered a weakness in American education, based primarily upon our lower scores compared to other developed nations. While standards are being raised nation-wide, this means very little if teachers are
ill-equipped to implement them. By researching how teachers use technological tools to solve informal inference tasks, we can gain insight into how these tasks might be implemented in classrooms, and we can be aware of misconceptions that may occur. Additionally, by observing what features of the data that teachers attend to and use as supporting evidence to form their conclusions, we can be mindful of potential ways that students may solve an informal inference task. Finally, by being cognizant of potential gaps of knowledge or subject matter that teachers have in education programs, we can make these programs stronger to prevent these pitfalls from occurring. Preventing the misconceptions of teachers will undoubtedly save future students from making the same mistakes.


