ABSTRACT

GONZALEZ, MARGGIE. Changes in Middle School Students' Ability to Engage in Informal Statistical Inference: A Probabilistic Approach. (Under the direction of Hollylynne Stohl Lee.)

Reasoning informally about statistical inference is being able to make inference about data without using any formal statistical method or procedure. One of the purposes of this investigation is to develop a framework to describe how students' informal statistical inference, in a probabilistic context, changes while being exposed to instruction. The author refers to probabilistic inference when talking about informal statistical inference in a probabilistic context. The framework is based on the levels of the SOLO Taxonomy (Biggs and Collis, 1982). A student in the pre-structural level makes generalizations based on their intuitions or previous experiences, shows no awareness of variability, does not consider sample size, and does not perceive sampling as a process to collect data. At the upper level of reasoning, students in the relational level make generalizations beyond the data collected using appropriate argumentation, consider the role of variability and sample size when making generalizations, and when possible propose an appropriate data collection process, including a description about how the results could be analyzed.

This study investigated middle school students as they were engaged solving several tasks in a 12-day instructional sequence in probability using Probability Explorer (Stohl, 1999-2005), a computer-based simulator. Tasks were selected across the instructional program with the objective of investigating how students' ability to engage in informal

statistical inference progressed from the beginning of the program to the end. To measure progress, responses are used from six students' responses on a selection of tasks from a pre-test, pre-interview, in-class worksheets, post-interview, post-test, and a retention test given two months after the program ended.

The findings indicate that most of the students are more aware of the effect of sample size when drawing conclusions and making probabilistic inferences at the end of the instructional program. They are also more prepared to investigate a situation, proposing a data collection process and a way to assess the results. In addition, all students demonstrated ability to reason at higher levels when working with a partner.

Changes in Middle School Students' Ability to Engage in Informal Statistical Inference: A Probabilistic Approach

by
Marggie Gonzalez

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

Mathematics Education

Raleigh, North Carolina

2010

APPROVED BY:

_____	_____
Hollylynne Stohl Lee, Ph.D. Committee Chair	Allison McCulloch, Ph.D.
_____	
Roger Woodard, Ph.D.	

DEDICATION

To my husband Alejo

BIOGRAPHY

Marggie D. Gonzalez-Toledo was born April 16, 1979 in Arecibo, Puerto Rico. She is the daughter of Ovidio Gonzalez and Ana Toledo and the oldest of three siblings. Marggie was raised in Utuado, a humble town in the mountains of Puerto Rico, where she attended elementary, middle, and high school. She graduated from Luis Muñoz Rivera High School with honors in 1997 and began a career in mathematics education at The University of Puerto Rico in Mayaguez, Puerto Rico during the same year. She graduated cum laude in May 2002 with a Bachelor in Science in Mathematics Education. After having a great experience in her student teaching Marggie decided she wanted to continue her professional career in the educational field, but for personal reasons she was not able to pursue her dreams at that time. Instead she decided to pursue a Master in Science in Mathematics with emphasis in Statistics, which she completed successfully in December 2005. After completing her MS she got married and started teaching at the University of Puerto Rico in Utuado. She taught in this institution for almost three years having the opportunity to teach from elementary algebra courses to Calculus I.

In 2007 Marggie moved to Washington D.C. where she started to work at the US Census Bureau as a Survey Statistician for the Economic Census of Puerto Rico and US Territories. A year later, she moved to Raleigh, NC to fulfill another chapter in her live, to continue graduate school and pursue a Ph.D. With this goal in mind she plans to pursue a Ph.D. in Mathematics Education after receiving her master's degree.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my husband Alejo. His support and encouragement has been constant throughout the process of completing this work. Alejo, thank you for being there when I needed you most and always helping me through those days when I thought this was only a dream.

I would like to thank my advisor, Dr. Hollylynne Lee, for all of her support and encouragement. Thanks for believing in my abilities to complete this work at the same time I was dealing with other things in life. Thanks for pushing me to do the best of my abilities; your feedback was so helpful. I feel so privileged to have the opportunity to work with you. Thank you for being more than my advisor and teacher when I really needed it. It is hard to get through difficult times in life when you are far away from family. I appreciate you being by my side through my academic and personal journey. Thank you, also, to my committee members Dr. Allison McCulloch and Dr. Roger Woodard. Your support and feedback during this process has been invaluable.

I would also like to thank my family in Puerto Rico. I especially want to thank my father, Ovidio Gonzalez, my mother, Maggie Toledo, and my two brothers, Ovimael and Joel, for encouraging me to pursue my goals and helping me through one of the hardest moments in my life. Even though they are far away, they are always there for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Quantitative information is everywhere. Interpreting data and being able to use it to make predictions and decisions are essential tools in order to be informed, educated, and competent citizens in our modern society. Individuals must have the ability to critique, interpret, evaluate, and express their own opinion about the information they receive (Shaughnessy, 2007). All citizens need to be able to critically read and evaluate tables, graphs, and news reports. They also need to critically interpret what is presented by the government, scientists, advertisements, and politicians.

The ability to critically reason with statistical ideas and make sense of statistical information is known as *statistical reasoning* and it involves making interpretations based on sets of data, representation of data, or statistical summaries of data (Garfield & Ben-Zvi, 2006). In order to be informed citizens, individuals need to reason statistically and make inferences that will help them make better decisions in their life. "Being able to properly evaluate evidence (data) and claims based on data is an important skill all students should learn as part of their educational programs" (Garfield & Ben-Zvi, 2006, p. 3) since they will need to informally infer conclusions from data presented through different representations (e.g., tables, graphs) by the media (e.g., news reporters, newspapers, magazines) in their everyday lives.

Shaughnessy quoted H. G. Wells as saying "statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write" (Shaughnessy, 2007, p.964). This quote appeared in the book *How to Lie with Statistics*, published in 1954 by Darrell Huff (Huff, 1954, p. 2). More than 30 years later the National Council of Teachers of Mathematics (NCTM, 1989) has suggested that teaching of data analysis and probability start in the early elementary levels, acknowledging the important role statistics and probability plays in the way we conducted our everyday life. By recommending the teaching of data analysis and probability the United States joined countries such as Australia, England, and the United Kingdom, who have had data analysis and probability in their respective mathematics curriculum documents (as noted by Aspinwall & Tarr, 2001; Jones, Langrall & Mooney, 2007; Pratt, 2005). Nowadays, in the United States, data analysis and probability figures as one of the five content strands of mathematics recommended by the NCTM (2000).

The data analysis and probability strand recommends that students from K to 12 should be able to collect data, organize it, use graphs to display the data, and learn methods to analyze the data collected, as well as learn basic concepts and applications of probability, with a special emphasis in making connections between probability and statistics (NCTM, 2000). In general, the NCTM (2000) has advocated that all students from K to 12 develop and evaluate inferences and predictions based on data. In particular, "upper elementary and early middle-grades students can begin to develop notions about statistical inference" (p. 50). The data analysis and probability strand also calls for students to have "basic understanding

of probability to make and test conjectures about the results of experiments and simulations" and to be able to construct sampling distributions (NCTM, 2000, p.324).

Since curricula documents were updated in the early 90's, an increasing interest for understanding how students learn statistics and probability has emerged in the United States and in other countries, and with it an increasing demand for research in the teaching and learning of statistics and probability (Jones et al., 2007; Shaughnessy, 2007). During the 1980's and early 1990's, research on students' understanding of probability and statistics was nascent and mostly concerned in studying students' misconceptions about probability and statistics (Shaughnessy, 1992). More recently, research has started to pay attention to students' understanding of statistics and students' statistical thinking (Shaughnessy, 2007). Students' informal statistical inference is a topic of current interest within the mathematics educators' community (Ben-Zvi, 2006).

In literature, several definitions have been presented to define informal statistical inference. Some of these definitions will be presented in Chapter 2, as well as a definition developed by the author of this investigation. In general, there are four key components researchers should look for when studying students' informal statistical inference. Students should be able to make predictions about the population from the samples collected, articulate data-based arguments to support their findings, incorporate probabilistic language, and use their prior statistical knowledge (Makar & Rubin, 2009; Zieffler, Garfield, delMas & Reading, 2008). An increased use of these key components is an indicator of growth in students' informal statistical inference.

In the *Principles and Standard for School Mathematics*, the NCTM (2000) encourages teachers to have students involved in situations in which sampling and simulations help them quantify the likelihood of an uncertain outcome (e.g., when a fair coin is flipped, when a regular six-sided die is tossed) and also to have students involved in situations where they have to compare the likelihood of events, both theoretically and experimentally. The idea that individual events are not predictable but that a pattern can be revealed and used to predict the likelihood of an outcome is an important concept that serves as the base for the study of statistical inference (NCTM, 2000, p. 50), and could serve as a place to link probability and statistics.

Accordingly with Pratt, Johnston-Wilder, Ainley and Mason (2008) informal statistical inference is concerned with identifying patterns in the "underlying" population. But recent research has been interested in how children use samples of data to think informally about populations (Ben-Zvi, 2006; Pfannkuch, 2006a). Research tasks typically are focus on having children make statements about a finite population from which the data has been drawn. Pratt and his colleagues (2008) have a different focus for their tasks. He stated that "other possibilities for informal inference exist" (p. 108) referring to situations where the population cannot be described in terms of totals, but instead through probability distributions.

The focus used in some of the tasks created by Stohl & Tarr (2002a) is similar to Pratt et al.'s (2008) focus, where totals were not given while students were engaged in tasks

involving probabilistic thinking. Instead of having students calculate probabilities of certain events, with all the totals given, some of the tasks challenged students to estimate proportions for certain events. Students used a simulation tool, ran an experiment as many times as they wanted, and made generalizations about the proportions using, for example, a bar graph which represent an empirical probability distribution to support their arguments. Through probability distributions, students can appreciate the importance of variability and sample size in statistical inference. In that sense, informal statistical inference can be introduced using situations involving probability where students can develop a strong conceptual base in which to build a more formal study of statistical inference (Paparistodemou & Meletiou-Mavrotheris, 2008). Students' understandings of distributions, variability, and sample size will be presented in Chapter 2.

Empirical probability distributions can be constructed by hand by asking students to collect the data, as Konold and Kazak (2008) did with their students. But empirical probability distributions can also be easily obtained using simulation tools. Simulations tools afford students access to relatively large samples that can be generated quickly and that can be modified easily (NCTM, 2000). Using computer simulations students can generate samples of different sizes and compare the relative frequency of an outcome (e. g., heads when flipping a coin). If relative frequency discrepancies are found between each sample students will be interested in investigating whether to keep their hypotheses or change it based on evidence obtained from the simulation (Lee, Angotti, & Tarr, 2010). Simulation tools are important tools in informal statistical inference because of the multiple

representations students can get in a short period of time and because large samples sizes can be generated quickly. Also students can have a visualization of the data as it is generated and accumulated dynamically in graphical displays. The visualization motivates students to notice variability when small and large samples are used (Lee, 2005). The availability of multiple representations helps students connect the sampling with relative frequencies and the probability distribution. Simulation tools also help students make the connections between experimental and theoretical probability (Pfannkuch, 2005; Stohl & Tarr, 2002). Some of the research that has been done using simulation tools and their primary findings will be discussed in Chapter 2.

Statement of the Problem

Since research in students' statistical inference is a topic of current interest, researchers have emphasized the need for more research in the area in order to gain knowledge about how students learn statistics, especially how they use data to infer characteristics of the underlying population (Jones et al., 2007; Paparistodemou & Meletiou-Mavrotheris, 2008). Specifically Shaughnessy (1992, 2007) and Jones et al. (2007) have called to the necessity of teaching experiments in probability in order to determine the impact instruction has on students' probabilistic thinking.

This research responds to that call and aims to contribute with additional results and knowledge about the effect instruction has in students' informal statistical inference. This study focused on investigating how students' abilities to make informal statistical inferences

changes during an instructional program. The changes were characterized following a framework the author developed based on the existent work on informal statistical inference in literature.

This study examines middle school students' informal statistical inference while engaged in a series of challenging tasks during a 12-day instructional program in probability. Students' informal statistical inference will be examined before, during, and after instruction using a sample of the tasks. Selected tasks from pre and post interviews, pre and post tests, and students' work on several instructional tasks were selected in order to create sequences of tasks in such a way that each sequence had tasks with similar objectives. This will allow an examination of how students' informal statistical inference is, or is not, changing as a result of the instruction.

Organization of the Paper

Chapter 2 will present a review of existing literature in order to illustrate the importance of having students exposed to informal statistical inference at an early stage in their lives. The literature gives a preamble to the investigation done by the author. The specific research questions investigated in this study are stated in Chapter 2.

Chapter 3 will describe the methodology used in the study, including the participants, data sources, and how data was analyzed to answer the research questions. Chapters 4 through 7 will present the results found during the study, and conclusions and implications about the results will be presented in Chapter 8.

CHAPTER 2

LITERATURE REVIEW

Acknowledging the impact instruction has in students' probabilistic reasoning is of great importance for teachers and researchers in the statistic education area. Investigation of how students' informal inference develops while exposed to instruction is of current interest in the mathematics education community. In the last two decades at least three exhaustive reviews have been prepared to describe what has been done and what is still needed regarding statistics education research (Jones et al. 2007; Shaughnessy, 1992, 2007). The literature review includes a discussion of what previous research has found about students' informal statistical inference and students' probabilistic reasoning as well as why is it important to consider the impact instruction has in that reasoning.

In order to enhance students' statistical reasoning and probabilistic thinking, it is important to know what key aspects need to be taken into account when teaching both topics. Teachers need to understand the importance of engaging their students in challenging tasks that enhance students' understanding about informal statistical inference and probabilistic reasoning. In addition, researchers need to acknowledge what types of tasks are adequate when conducting research in order to study students' informal statistical inference. As indicated by Zieffler et al. (2008), researchers need well-designed tasks that allow them to capture and evaluate students' reasoning. These tasks must provide students with

opportunities to reason about a situation, make conjectures, and use data to support these conjectures.

The next section includes a discussion about what informal statistical inference is and what previous research has found about its learning. Following this is a discussion on students' statistical inference in a probabilistic environment and a review of how students learn two of the most important concepts in statistics and probability: sampling and variability. Next the author will present several frameworks that have been developed in the last few years trying to accommodate students' thinking and teachers' pedagogical issues in the areas of informal statistical inference and probability inference, including the description of the framework the author developed during this investigation. The research questions will be presented at the end of this chapter.

Informal Statistical Inference

Before students are exposed to informal statistical inference at the secondary level, they are usually exposed to Exploratory Data Analysis (EDA) at the primary level (Tukey, 1977). Exploratory Data Analysis (EDA) was developed by John Tukey and it mostly emphasizes ways to uncover, display, and describe patterns in data (Ben-Zvi, 2006). Tukey (1977) defined EDA as "the examination of data with minimal preconceptions about its structure through which it is hoped that relationships and patterns will be uncovered" (p. 151). Its goal is to make sense of data with a heavy reliance on visual displays as analytical tools. The *Principles and Standards for School Mathematics* (NCTM, 2000) calls for

students to collect, analyze, and display relevant data starting in kindergarten. Moreover, students in the upper-elementary level should be able to collect data using observations and experiments and display it using tables and graphs (NCTM, 2000, p. 178). Students should explore with data in their primary levels, and be ready to make statistical inference in an informal way both at the primary and secondary levels. Pedagogically, EDA can be used to give students opportunities for open-ended data explorations aimed to develop a sense of basic concepts of statistics.

What is informal statistical inference?

Many authors have described informal statistical inference in different ways. Rubin, Hammerman, and Konold (2006) described it as a reasoning that involves consideration of properties of aggregate including signal and noise and types of variability, the effect of sample size on the accuracy of population estimates, and controlling for bias in order to get a sample that is representative of the population. They also mentioned the importance of students recognizing whether an event is impossible, unlikely, likely, or certain (Hammerman & Konold, 2006). Ben-Zvi (2006) stated that argumentation and data-based evidence are two important aspects of informal statistical inference. Zieffler et al. (2008) agreed with Ben-Zvi and defined it as the way students use argumentation and data-based evidence to support the connections they build between observed sample data and the unknown population. Pratt (2005) and Pfannkuch (2005) agreed in their definition stating that statistical inference is concerned with drawing conclusions about specific characteristics of a particular population based on evidence obtained by a sample.

Rubin et al. (2006) and others have found that students have some intuitions about data which can be refined towards reasoning that has inferential qualities. In that sense the authors define informal statistical inference as making generalizations beyond the data, using proper language, argumentation, and data-based evidence to support their findings. The inference is informal in the sense students have not had any formal instruction of methods of statistical inference, but use their previous knowledge or previous experiences to come up with answers of questions about an underlying population.

Pedagogical frameworks on informal statistical inference

Pfannkuch (2005) presented a pedagogical framework which focused on making teachers aware of the reasoning students need to experience and develop for inference in four main areas: measures of center, distributions, sampling, and drawing conclusions based on informal inference. This framework was a result of the analysis of Grade 10 students' comparison of box plots using temperature data from two cities. Students in the study had to formulate a question, analyze the data, draw a conclusion and justify it.

Pfannkuch (2006a) conducted a teaching experiment to analyze teachers' reasoning when comparing box plots. From the data collected during the study she identified seven elements important for statistical reasoning: hypothesis generation, summary, shift, signal, spread, sampling, explanatory/context, and individual cases. She used these seven elements and proposed a descriptive model of reasoning with box plots. Watson (2008) used an adapted version of Pfannkuch model with Grade 7 students using *TinkerPlots*. She adapted

the framework to be used to analyze students' comparisons of distributions using hat plots. Pfannkuch (2006a) and Watson (2008) used the SOLO Taxonomy in their analysis of students' comparison of distributions using box plots and hat plots, respectively. The SOLO Taxonomy will be described later in this chapter.

Following Pfannkuch's (2005, 2006a) framework, it is important for teachers to help their students recognize the presence of signal and noise in any collected data. Even though individual events cannot be predicted, the aggregate data follows a pattern which makes possible to make predictions about characteristics of the group. Using the appropriate learning experiences such as those suggested by Konold and Pollatsek (2002) could lead students to have a clear understanding about signal and noise (Pfannkuch, 2005). For example, when comparing the reading scores on the National Assessment of Educational Progress (NAEP) there are other variables to have into account before saying that the mean score increases from one year to the other (Konold & Pollatsek, 2002). Students need to learn to think about data as a mixture of signal and noise and to recognize the effect signal and noise have in the interpretation of data (Konold & Pollatsek, 2002).

The role of teachers

The role that teachers play on the instruction of informal statistical inference has an enormous impact on students' understanding and learning. Researchers recommend different tasks teachers can implement in their classroom in order to foster students' informal

statistical inference (Zieffler et al, 2008). But there are other aspects teachers need to consider when designing and implementing a task. Cobb and McClain (2004) discussed some principles teachers should take into account when designing tasks driven to help students develop statistical reasoning.

- *First*, the task should be focused on developing only one central statistical idea instead of presenting a set of tools and procedures.
- *Second*, the teacher should use real and motivating data sets to engage students in making conjectures that are meaningful to them.
- *Third*, the teacher should determine the structure of the classroom activities more appropriate to help students develop their statistical reasoning (i.e. group vs. individual activities, whole class discussion).
- *Fourth*, the teacher needs to determine the appropriate technology tool that will allow students to explore the data and test their conjectures, and
- *Fifth* the teacher should promote classroom discourse focused on statistical arguments and significant statistical ideas.

Garfield and Ben-Zvi (2008) used these principles as a guide in two courses, one for graduate students pursuing a master in elementary mathematics education in Israel and one for in-service secondary mathematics teachers in the United States. Garfield and Ben-Zvi mentioned that preparing teachers to teach statistics is challenging and that those teachers need to be prepared to create a learning environment to develop a deep and meaningful

understanding of statistics in their students.

Literature on informal statistical inference

A wide selection of studies have been conducted in order to study students' informal statistical inference (Ben-Zvi, 2006; Pfannkuck, 2006a, 2006b; Rubin et al, 2006;Watson, 2001, 2008; Watson & Donne, 2008, 2009). Research has found that even young children can collect data, display it in an adequate way, and make inferences about the underlying population based on the data collected. Moreover, in some cases, students incorporate appropriate language in their arguments (Paparistodemou & Meletiou-Mavrotheris, 2008). Most of these studies have had students using dynamic software such as *Fathom* (Finzer, 2007) and *TinkerPlots* (Konold & Miller, 2004) to analyze data.

Researchers agree that exposing students to exploratory data analysis in early grades could help students develop statistical thinking (Ben-Zvi, 2004; Tukey, 1982). NCTM (2000) suggests that K-2 students should be able to formulate questions, collect data, and display relevant data in graph such as a bar graph or a line-plot graph (p. 109). Statistical ideas such as aggregate, measures of central tendency, sampling, and variability could be introduced in a very informal way using data that has some meaning for students. In middle grades students should be exposed to statistical inference but without the formal methods used in a statistical inference course. In doing so, students would develop a sense of what it means to have a random sample and how variability is involved in any statistical situation, two concepts that are the primary assumptions in a statistical situation.

Distributions, variability, and sampling

According to Cobb and McClain (2004) teachers should focus their lessons on developing central statistical ideas (i.e., concepts) rather than presenting disconnected tools and procedures. Distributions, variability, and sampling are three of these central statistical ideas along with center and randomness (Garfield & Ben-Zvi, 2008).

Variability stands in the heart of statistics theory and practice (Ben-Zvi, 2004) and is a fundamental component of statistical thinking (Garfield & Ben-Zvi, 2005); also researchers have found that its understanding is a challenge for teachers and students (Hammerman & Rubin, 2004). Variability exists between and among samples (distributions), locating distribution in the heart of statistics. Distributions reveal important information about the center, shape and spread of the data (Garfield & Ben-Zvi, 2008).

Sampling is also a central statistical idea in statistics. Samples are randomly drawn and used to make predictions about certain characteristics about the population from which the data was drawn. Recognizing the importance of sampling and sample size is one of the biggest problems related with informal inference and with it the interconnection between sample size and variability (Pfannkuch, 2005).

According to Garfield and Ben-Zvi (2005) there are several areas of knowledge of variability, which can be "viewed as building blocks for constructing deep understanding of the complex concept of variability" (p. 93). The first aspect, *developing intuitive ideas of variability*, includes acknowledging that variability is everywhere in statistics and

understanding the different reasons and sources of variability. A second important aspect is to understand *how variability is described and represented*. Using multiple representations offer different information about variability in data. Using a pie graph or a bar graph could give students different information about variability in data. Another aspect is *variability in random events*. As indicated before, samples and distributions vary and this variation could be reduced as the sample size increase. To make appropriate statistical inferences from data generated by random events students should recognize the important role sample size plays.

Variability is also important to make *comparisons*. Variability exists within samples and distributions as well as across samples and distributions. In order for students to start noticing variability across different samples obtained from the same population they need to look at the data as a whole instead of at individual cases. According to Hammerman and Rubin (2004) the use of visualization tools help students move from looking at individual cases (additive reasoning) to look at the data as a whole (multiplicative reasoning), whereas Ben-Zvi (2006) mentioned that the use of *TinkerPlots* focuses students attention to the individual cases and hinders spontaneous progress to aggregate views of the data. Hammerman and Rubin (2004) and Ben-Zvi (2006) studies showed contradicting results, indicating that whether or not students moved from looking at individual cases to look at the data as a whole will depend on how appropriately the activities and the use of the technology tool are. Teachers need to create activities that focus student's attention to the data as a whole by mean of, for example, direct questioning.

Ben-Zvi (2004) has studied, and identified, several components that could be use to evaluate students' thinking and understanding about variability. Those components are acknowledging, measuring, explaining, controlling, describing, and representing variability. Students should be able to *acknowledge* variability and *describe* it; they should *explain* variability in the context of the problem and recognize where that variability comes from. Also students should think of ways of *controlling* some aspects of the experiment in order to reduce the variability. Students' conception of variability could be measured in different statistical context, such as reasoning about distributions, dealing with samples and sampling, reasoning about the outcomes of a probability experiment, and comparing data sets.

The Mus-Brush Company produces mushroom brushes, using a large machine whose output is on average 215 brushes every two minutes *if it is working normally*.

If the electricity to the machine is interrupted, even for a brief time, it will slow down such that the output of the machine will be 10% lower on average.

The Mus-Brush Company was robbed last night; in forcing the door open, the thief disrupted the electricity and the machine became less productive from that time on.

There is a prime suspect who has an alibi between midnight and 3AM (he was seen at a bar), so the police have a special interest in determining if the break-in occurred before midnight or after 3, since the suspect has no alibi for those time intervals.

We have data on Mus-Brush production every two minutes from 8PM until 5AM. Our job is to decide whether there is enough evidence to argue that the break-in occurred between 12 and 3, thus getting the suspect off the hook.

Figure 2.1. The Mus-Brush Company Task (from Rubin et al., 2006)

Rubin et al. (2006) and Pfannkuch (2006a, 2006b) described how teachers make inferences, and consider variability, in different scenarios. Rubin et al. analyzed teachers reasoning when working with the Mus-Brush Company task (Figure 2.1) using *TinkerPlots* whereas Pfannkuch had teachers introducing and discussing box plots to her students in a

non-technological environment. Teachers working on the Mus-Brush Company task had to give argumentation based on the data in order to help to solve the given situation. In general the teachers seem to have no problem understanding variability in the context of the problem. On the other hand, Pfannkuch found that teachers have difficulties articulating the difference between variability within and between box plots. Pfannkuch used the data gathered from the study to propose a descriptive model of reasoning from box plots.

While Rubin et al. (2006) and Pfannkuch (2006a, 2006b) studied teachers' statistical reasoning, Ben-Zvi (2006), Paparistodemou and Meletiou-Mavrotheris (2008), and Watson and Donne (2008) studied young students as they engaged in tasks using the dynamic software *TinkerPlots*. They found that young children can draw conclusions based on data and make inferences about a larger population. Paparistodemou and Meletiou-Mavrotheris also found that students can incorporate expressions about uncertainty such as 'more likely' and 'might be'. These expressions including uncertainty are a key component in students' informal statistical inferences, as indicated by Makar and Rubin (2009).

Ben-Zvi (2006) had 75 fifth grade students collected their own data using a 19-item questionnaire. Some of the questions were gender, age, body measurements, home to school distance and time, etc. Each student in the class had to randomly choose three students in grades 2, 4 and 6 and collect data from them, as well as themselves. They also needed to enter the data to *TinkerPlots*. Then students, working in pairs, were asked to choose a question they found interesting and investigate it using a small sample (n=8). For this

investigation, students were focused on questions about relationship between variables. Once the students had investigated their questions using the small sample they were asked whether their hypothesis would hold in larger samples (n=16). The sample size was increased to the whole class (n=80) and at the final stage of the study students compared a sample from their school (n=240) with a sample from all UK students (n=200). His results showed that the design of the learning trajectory based on growing samples (increasing the sample size from n=8 to n=240) helped students to improve their statistical reasoning. They observed progress from additive to multiplicative reasoning, consideration of aggregate views of data, acknowledge of the importance of larger samples, and accounting of variability. They also mentioned students showed great interest in the task and were engaged in the investigation because students felt represented in the data. As Cobb and McClain (2004) mentioned, teachers should use real and motivating data sets to engage students in making conjectures that are meaningful to them.

Paparistodemou and Meletiou-Mavrotheris (2008) had students collecting their own data and explore it using *TinkerPlots*. They created a 16-item questionnaire about nutritional habits and administered it to a total of 120 students in the school. Similar to Ben-Zvi (2006) students collected their own data and entered it into *TinkerPlots*. The task was aimed to foster students' ability to collect and represent data, as well as students' ability to propose and justify predictions based on data. In the study, Paparistodemou and Meletiou-Mavrotheris assessed students' growth in understanding and reasoning about statistical inferences. They found that statistical instruction could develop students' informal statistical inference at an

early age (third grade). Children in the study were able to draw conclusions based on the data collected (e. g., "most of the children who are playing with scissors belong to Grade A"), draw inferences about a larger populations (e. g., "most of the students at the school are exercising"), and draw inferences about an unknown population expressing uncertainty (e. g., "it is more likely for boys not to eat any sweet"). Expressing uncertainty when drawing inferences is an important element to consider when studying students' informal inference (Makar & Rubin, 2009).

Watson and Donne (2008) had students investigating hand reaction times. Similar to the previous studies, the authors had students collect their own data and enter it to *TinkerPlots*. Whereas Ben-Zvi's (2006) based his learning trajectory on a growing samples sequence, Watson and Donne introduced populations first rather than later in the investigation process. They found little is gained by introducing populations first. The authors used the descriptive model developed by Pfannkuch (2006a) and used the SOLO Taxonomy (Biggs & Collis, 1982) to categorize the relationships among the constructs in the framework, similar to Pfannkuch's framework. As most of the students moved forward in the lessons, their SOLO levels were also increasing showing students were starting to put ideas together in order to reach conclusions. The SOLO Taxonomy has been widely used in mathematics education and will be described in more detail later.

Ben-Zvi and Amir (2003) engaged three second grade students in a non-technological exploratory data analysis activity about losing milk teeth. The primary goal of the study was

to find how young students come to reason about distributions and what their intuitive emerging conceptions of distributions are. The children were engaged in a task where they needed to conjecture about the expected number of teeth students lose from kindergartners to fourth grade. They found students either constructed a *flat distribution*, which means they included all possible outcomes similar to a sample space, or a *distribution sense* emerged, meaning that the student include different possible outcomes and how often he expected those outcomes to appear (similar to sample space and density). But most of the students were consistent using the *flat distribution*, and only one student showed *distribution sense*.

Informal Statistical Inference in a Probabilistic Context

Previous sections were focus on understanding how students reason about statistical inference in an informal way and what are the important aspects to consider when analyzing students inference in statistics. In this section the author will discuss why it is important to engage students in informal statistical inference in a probabilistic context in first place. Hereafter the author will use *probabilistic inference* to refer to informal statistical inference in a probabilistic context.

The research presented in previous sections described students engaged in different scenarios where the main focus was to make informal statistical inference about a specific characteristic of a population based on a collected sample. The inference involved to make generalizations about parameters such as the typical height for a third grader and the typical number of teeth students from kindergartners to fourth grade loses (Ben-Zvi & Amir, 2003).

Probability was developed in response to situations about games of chance (Pfannkuch, 2005); for example, to estimate the chance of winning the lottery, to predict the weather, or to estimate the likelihood of a toothpaste cap to lie upside down.

Probabilistic inference is also based on data collected and the main objective is to predict the underlying probability distributions, obtained theoretical or empirically through experimentation or simulations. Probabilistic inference has been widely used with students in contexts where the theoretical probability distribution is known (i.e., coins, dice, and spinners) (Jones, Langrall, Thornton, and Mogill, 1997); but how do students reason about situations where the theoretical probability distribution is not known? According to NCTM, "through the grades, students should be able to move from situation for which the probability of an event can readily be determined to situation in which sampling and simulations help them quantify the likelihood of an uncertain event" (p. 51). Engaging students in the later could be used to introduce important statistical ideas such as distributions, variability, and sampling. Probabilistic inference can be use as a preamble to informal statistical inference, starting with deterministic situations (theoretical probability is known), and then moving to non-deterministic situation (theoretical probability is unknown) where sampling and simulations can be use to quantify the likelihood of an uncertain event.

Probability can be estimated using theoretical analysis (classical approach) or based on empirical data (frequentist approach). Jones et al. (2007) found that both the classical and the frequentist approaches were included in mathematics curricula around the globe since

early 90's. This frequentist approach, grounded in the empirical law of large numbers, has made its appearance recently into school curricula. The empirical law of large numbers was formulated by Jacob Bernoulli and states that the probability of a large difference between the relative frequency of an outcome and the theoretical-derived probability limits to zero as more trials are collected (Sedlmeier & Gigerenzer, 1997). Since the frequentist approach is nascent, more research is needed in order to study students' conceptions of the law of large numbers; students' conceptions of experimental probability, its connection with theoretical probability, and the importance of sample size (Jones et al., 2007).

Literature in probabilistic reasoning

Although a wide selection of research has been done to study students' statistical informal inference (e.g., Ben-Zvi, 2006; Paparistodemou & Meletiou-Mavrotheris, 2008; Rubin et al., 2006; Watson & Donne, 2008/2009; Watson 2001/2008), a narrow selection have examined the growth of students' probabilistic reasoning while being exposed to instruction (Aspinwall & Tarr, 2001; Berenson, 1999; Fischbein & Schnarch, 1997; Jones, Langrall, Thornton & Mogill, 1999), and only a few have used technology as a tool for students to carry out simulations (Lee, Angotti, & Tarr, 2010; Pratt 1998/2000; Pratt, Johnston-Wilder, Ainley, & Mason, 2008; Stohl & Tarr, 2002a/2002b; Tarr, Lee, and Rider, 2006). Exploratory data analysis tools such as *TinkerPlots* (Konold & Miller, 2005) and *Fathom* (Finzer, 2007) have been used to study how students use sample data to generate hypothesis and make generalizations based on data. Other tools such as *ChanceMaker* (Pratt, 2005), *InferenceMaker* (Pratt et al., 2008) and *Probability Explorer* (Stohl, 1999-2002) have

been used with positive results to study students' probabilistic reasoning (refer to next section).

Shaughnessy (1992) expressed the need for research on teaching and learning of probability, specifically the need of teaching experiments in probability in order to determine the impact instruction has on students' probabilistic thinking. Nowadays, although mathematics educators have been more involved in the research of statistics and probability, studies that examine the effect instruction has in students' statistical thinking are still needed (Jones et al., 2007). Moreover, Jones et al. called for research to examine the effect instruction has in students' conceptions of experimental probability and its relationship with theoretical probability. It is important to acknowledge what type of instruction is beneficial for students to develop probabilistic reasoning and what type of task are more appropriate to reveal students thinking in order for teachers and researchers to be able to construct appropriate assessments and curricular materials to foster students' statistical informal inference.

The National Council for Teachers in Mathematics recommended students in secondary levels to carry out simulations of random phenomena, to interpret results obtained from these simulations, and to make inference about the underlying population based on data (NCTM, 2000). Even though the NCTM has given an especial importance to experimental probability, theoretical probability and the role sample size plays in that relationship, Jones et al. (2007) found that little research has been conducted in that specific area. Studies have

addressed how students are able to recognize the importance of large sample size when making inference using experimental probability (Aspinwall & Tarr, 2001; Stohl & Tarr, 2002a).

Aspinwall and Tarr (2001) study was focused on students' understanding of experimental probability as it relates to sample size. In their experiment, Aspinwall and Tarr used a task simulating a car race. Students were asked to pick the car they thought will win the race. The car will move one step forward in the race every time the car's number showed up as the sum of the two dice. At first, students viewed the result of simulation as irrelevant and used subjective reasoning to determine the number of trial (the sample size). They found students typically believed that any sample, small or large, should reflect the parent distribution. Their results indicate that although middle school students are usually unaware of the relationship between experimental probability and sample size, using activities that simulate random phenomena can foster conceptual development. They also found that engaging students in simulations does not always promote growth in their understanding of probabilistic concepts. However, simulations did help some students developed a better understanding of empirical probability.

Berenson (1999) interested in study students representations of probability, especially sample space, used game-playing situations to engage 16 eight grade students in decision making and testing. Students played games following predetermined rules for scoring point. If after playing the game they thought the game was not fair they needed to redesign the

scoring of the game, test it again and redesign if still being unfair. First, students played the Once Dice Game, where player A scored a point every time the dice rolled a 1, 2, 3, or 4 and player B received a point when the dice rolled a 5 or a 6. Then they played the Two Dice Game where each player received point accordingly with the sum of both die. Player A received points for sums 2, 3, 4, 10, 11, or 12 and player B for sums 5, 6, 7, 8, or 9. Results indicated students uses several ways to represent chance events. Some of them used *equal number* (EN) where they believe each player should have an equal number of die faces or an equal number of sums. Other though that certain sums had *more combinations* (MC) than others thus it was easier for the player to win. Probability simulations elicited student thinking in determining the fairness of several dice games.

In another study conducted by Stohl and Tarr (2002a) students were challenged to determine whether or not a dice was fair. They used the Schoolopoly task (Figure 2.2), which is intended for students to investigate whether or not dice sold by several companies are fair. Students needed to conduct a simulation, collect the data, analyze it, and interpret it. At the end students were asked to present their results to the class. One of the case studies analyzed by Stohl and Tarr started by saying the die was fair because "every single time doesn't have to be even, it's the luck" (p. 332), but later during their investigation they started to care about larger samples and to notice the die was, after all, unfair. Their results showed that after appropriate instruction students understand the interplay between empirical and theoretical probability, and recognized the importance of drawing samples with a large number of trials in order to make better inferences. This type of task, such as the Schoolopoly

task, "afford all students access to a real-world problem that requires them to collect and analyze empirical data and then formulate and evaluate data-based arguments" (Tarr, Lee, & Rider, 2006, p. 148). This type of task can "challenge students to make judgments and predictions about a population without the use of formal statistical methodology" (Zieffler et al., 2008, p. 51-52).

Schoolopoly: Is the die fair or biased?

Background
Suppose your school is planning to create a board game modeled on the classic game of *Monopoly*. The game is to be called *Schoolopoly* and, like *Monolopy*, will be played with dice. Because many copies of the game expect to be sold, companies are competing for the contract to supply dice for *Schoolopoly*. Some companies have been accused of making poor quality dice and these are to be avoided since players must believe the dice they are using are actually "fair". Each company has provided dice for analysis and you will be assigned one company to investigate:

Luckytown Dice Company	Dice, Dice, Baby!
Dice R' Us	Pips and Dots
High Rollers, Inc.	Slice 'n' Dice

Your Assignment
Working with a partner, investigate whether the dice sent to you by the company are fair or *biased*. That is, collect data to infer whether all six outcomes are equally likely and answer the following questions:

1. Do you believe the dice you tested are fair or biased? Would you recommend that dice be purchased from the company you investigated?
2. What *compelling evidence* do you have that the dice you tested are fair or unbiased?
3. Use your data to estimate the probability of each outcome, 1-6 of the dice you tested.

Collect data about the dice supplied to you. Note that each single trial represents the outcome of one roll of a "new" virtual die provided by the company.

Copy any graph and screen shots you want to use as evidence and print them for your poster. Give a presentation pointing out the highlights of your group's poster.

Figure 2.2. The Schoolopoly Task (Stohl & Tarr, 2002a)

Lee, Angotti, and Tarr (2010) reported results from those students reported by Lee and Tarr (2002a). They were focused on how the students investigate the fairness of a die and

how they use empirical data to support or refute a model of a die that assumes equiprobable outcomes. They used a framework that described the process of statistical investigation where they compared student's expectations and observations within an informal hypothetical testing cycle. Since the students had a previous experience with a dice their null hypothesis was the die was fair. Students collected data and used it to decide whether or not the current data set confirm or contradict their initial hypothesis. They found students developed notions about the importance of sample size and variability. Also the three pairs in the study seemed to reason about variability within a data set rather than across samples. They stated that "the dynamic nature of the technology may have contributed to this phenomenon." (p. 20)

In other studies, researchers have also investigated how students reason about small and large samples and even how students reason when selecting a sample size (Lee, 2005; Pratt, 2000). Lee (2005) described a series of strategies that she has observed in several studies, including her own work. If the students, for example, have a bag with N=10 marbles they will often use a strategy that include only n=10 trials and where the empirical results are taken from granted to represent the theoretical distribution. Lee named this strategy the "total weight approach" where n=N. This strategy often happened when the students know the theoretical probability beforehand. On the other hand, if students are engaged in a task where the theoretical distribution is unknown they start moving from the "total weight approach" to run a large number of trials. In doing so, students started noticing patterns in the samples with large number of trials. Lee named this second strategy the "evening out" phenomenon.

When the theoretical distribution is not known beforehand, students need to make statistical inferences from the data collected. In this sense the "evening out" phenomenon will help students understand the importance of sample size in statistical inference.

Use of Technology in teaching and learning probability

The research presented along this chapter has shown that the use of technology could improve students' conceptual understanding of important statistical ideas, such as distributions, variability, and sample size (Aspinwall & Tarr, 2001; Ben-Zvi, 2006; Hammerman & Rubin, 2004; Lee, Angotti, & Tarr, 2010; Pratt, 2005; Stohl & Tarr, 2002a). The technological tools used trough those studies are the *ChanceMaker* (Pratt, 2005), the *InferenceMaker* (Pratt et al., 2008), and *Probability Explorer* (Stohl, 1999-2002).

Pratt (2005) suggested the use of *ChanceMaker* in order to promote students reasoning about chance in stochastic situations. According to his research, children need to be challenge in order to appreciate their perspective of equiprobable situations. His main objective was to "gain fresh insights into how children's stochastic thinking evolved through the use of the [*ChanceMaker*]" (p. 179). The *ChanceMaker* consisted of gadgets (e.g., coin, spinner, and dice), whose behavior was controlled through a "working box." The working box was a representation of the distribution of the gadget. Students were allowed to explore the gadgets using tools such as pie charts to display results and to edit the content of the working box. By controlling the content of the working box, students were creating a probability distribution and generating data from it. He found the design of *ChanceMaker*

support students understanding of the importance of sample size in order to explain the behavior of certain gadgets (Figure 2.3).

Figure 2.3. The *ChanceMaker* (from Pratt et al., 2008, p. 110)

Some conclusions from Pratt's research are: (1) the tasks must provide evidence for the children to appreciate the power of the new ideas with their own intuitions; (2) the tasks should encourage students to decide about the number of trials by themselves in order for them to realize the problems with using small number of trials; and (3) different context should be use in order for students to appreciate the wide applicability of the mathematical idea (e.g., randomness).

Pratt, Johnston-Wilder, Ainley, and Mason (2008) presented results after a small group of children used the *InferenceMaker*, which work similar to the *ChanceMaker*. Within the design of *InferenceMaker* students were not able to see the working box and with that students were not aware of the underlying probability distribution (Figure 2.4). With *InferenceMaker* the user edit the working box and then hide it so that the students were

challenged to make inferences about the configuration of the gadget (e.g., coin, or dice) by generating data and charts. Their data indicates that their students did not recognize the importance of sample size in getting a reliable image of the probability distribution of, in this particular case, the dice. It is worth to mention that Pratt and his colleagues conducted clinical interviews and their results are not part of a teaching experiment. It might be possible to obtain different results if the software is used during instruction.

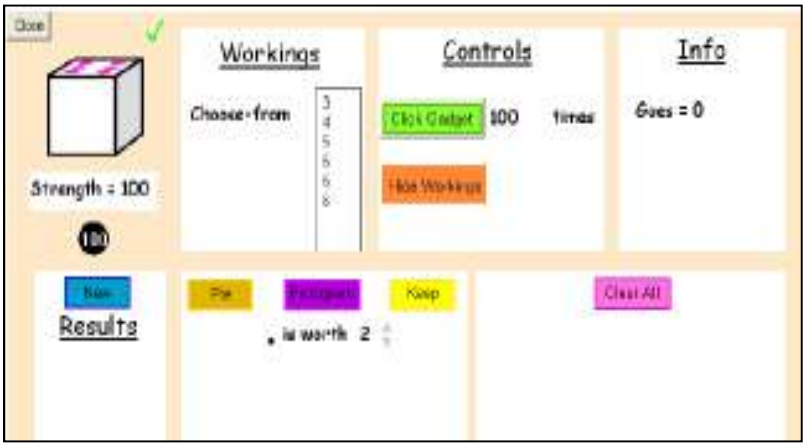

Figure 2.4. The *InferenceMaker* (from Pratt et al., 2008, p. 114)

Probability Explorer is a simulation tool developed by Stohl (1999-2002) where students can conduct experiments on a series of contexts (Figure 2.5). Similar to *InferenceMaker* (Pratt et al., 2008), *Probability Explorer* has a weigh tool that teachers can hide so that the students use the data collected to infer the probability distribution (Figure 2.6). With PE it is also possible to carry out long term experiments as it is possible to run up to 1000 trials. Stohl and colleagues have conducted teaching experiments using Probability explorer and result have been reported in Lee and Tarr (2002a, 2002b), Tarr et al. (2006), and

Lee et al. (2010). In general their results have shown that after appropriate instruction students understand the interplay between empirical and theoretical probability and develop notions about the importance of sample size and variability.

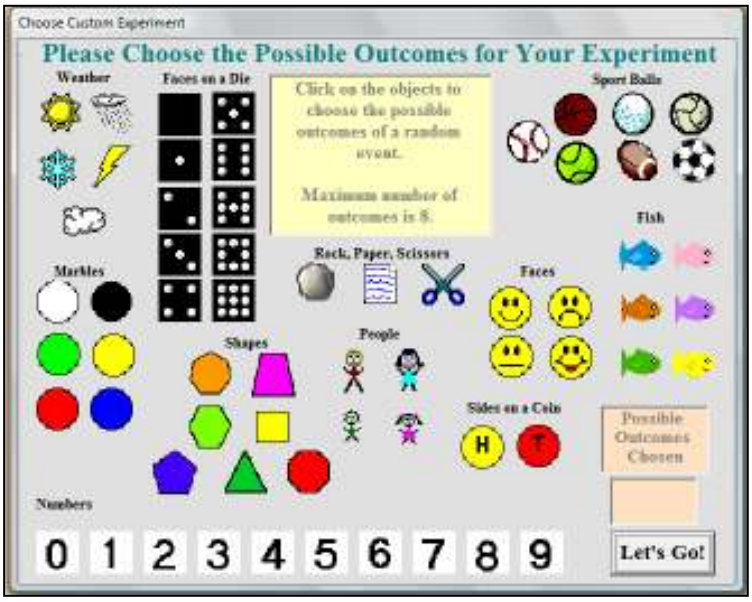

Figure 2.5. Design your own experiment in *Probability Explorer*

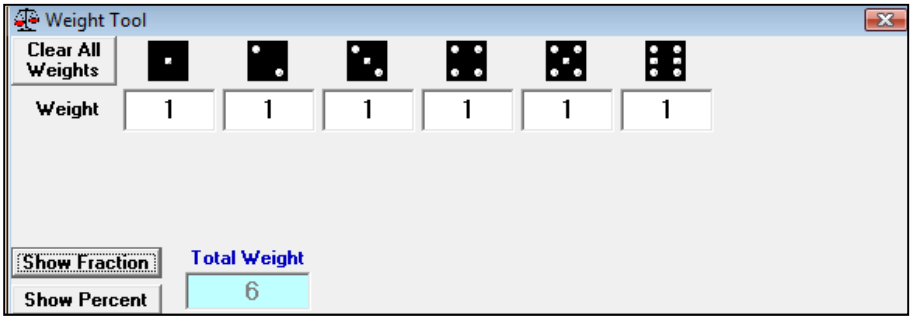

Figure 2.6. The Weight Tool in *Probability Explorer*

According to Chance, Ben-Zvi, Garfield, and Medina (2007), "technology has a great

potential to enhance students achievement [and] it will most likely continue to impact the practice and the teaching of statistics in many ways" (p. 3). However, they emphasized that in order for it to have an impact in education, technology needs to be used appropriately. When students use computer simulations they can generate samples and create probability distributions in an easy way, and they can make predictions and articulate arguments based on the data they generated. It is important to keep in mind that in order for simulations to be meaningful in statistical inference, teachers need to help students understand what the simulation data represent and how they relate to the problem situation (NCTM, 2000).

Frameworks in Literature

There are quite a few frameworks in statistics education and they are focused on different aspects. Several researchers have developed frameworks in order to think about key components researchers and teachers should look for when studying students' informal statistical inference (Makar & Rubin, 2009; Zieffler et al., 2008). Others researchers have developed frameworks to study students' probabilistic thinking (Jones et al., 1997) and students' inferential statistical inference (Watson, 2001; Vallecillos & Moreno, 2002. A description of the frameworks available through the literature to study informal statistical inference as well as probabilistic inference is presented next, followed by a description of the framework developed by the author of this investigation.

Frameworks to study informal statistical inference

Makar and Rubin (2009) and Zieffler et al. (2008) have developed frameworks in order to think about key components researchers and teachers should look for when studying students' informal statistical inference. Both agreed that two key components in such a framework are that students should be able to make judgments or predictions about the population based on the samples collected and articulate data-based arguments in order to support their claims. A third component in Zieffler's et al. framework is students' integration of prior statistical knowledge while Makar and Rubin added as a third component students' uses of probabilistic language. Both of these frameworks have been proposed based on theoretical reflection and have not been used yet by the authors to study students' informal statistical inference.

Makar and Rubin (2009) reported data from the first phase of an ongoing four-year study investigating the processes of teachers' learning to teach mathematics and statistics through inquiry in a problem-based environment. The authors used their framework to study four primary school teachers' teaching of inferential reasoning. Especially how those teachers encourage students to reason informally about statistical inference giving importance to the three components included in the framework. Their framework, however, has been never used to analyze students' informal statistical inference. Zieffler et al. (2008) presented their framework theoretically and included what types of task they think can be used to study informal statistical inference, but it has not been validated and/or used in any study with students. Among the tasks they selected are the Schoolopoly Task used in Stohl

and Tarr (2002a) and in Tarr et al. (2006) and the Mus-Brush Company Task used in Rubin et al. (2006).

The Guideline for Assessment and Instruction in Statistic Education (GAISE) report is a Pre-K-12 curriculum framework founded on the Principles and Standards for School Mathematics (NCTM, 2000). It was developed by Franklin, Kader, Mewborn, Moreno, Peck, Perry, and Scheaffer and endorsed by the American Statistical Association in 2005. The authors of the framework visualize statistical problem solving as an investigation process that involves four components: formulate questions, collect data, analyze data, and interpret results; and it is focused on the role variability plays in the statistical problem-solving process. The role of variability changes at each different component; variability must be *anticipated* when formulating a question, *acknowledged* when collecting the data, *accounted* when analyzing the data, and *allowed* when interpreting the results. The framework consists of three developmental levels that are not related to age, but rather are based on development in statistical literacy. The levels are consecutive one to the other, which means that every student must begin with Level A concepts and activities before moving to Level B, and must have the experiences of Level B before engaging in Level C concepts and activities. The authors view statistics education "as a developmental process" and proposed a framework for statistic education over three years. Again the levels are not tied to grade levels but to development of statistics literacy.

Vallecillos and Moreno (2002) developed the Inferential Statistical Thinking

framework in order to characterize and assessed the learning of basic statistics. Their framework consisted of four construct and four levels of thinking within each construct, and was based on the general cognitive model developed by Biggs and Collis (1982). Vallecillos and Moreno's four construct are: population and samples and their relationship, inferential process, sample sizes, and sampling types and biases.

Frameworks to study probabilistic inference

Contrary to the study of students' reasoning about statistical inference there are just a few frameworks to study how students reason about probabilistic inference. In this section the author will describe some of the frameworks that are available to study students understanding of probability.

Watson, Collis, and Moritz (1997) developed a framework to describe students' understanding of chance measurements. They administered a questionnaire to 1014 student in Grades 3, 6 and 9 in Tasmania. Only three items were analyzed in order to characterize students' understanding of chance measurement. The tasks included comparing the likelihood of individual events in situations like tossing a six-sided die, drawing names from a hat, and comparing distributions. Their framework was based on the SOLO Taxonomy. They proposed a model with two Unistructural-Multistructural-Relational hierarchical cycles. Their results showed that students in higher grades have more sophisticated thinking.

Jones, Langrall, Thornton, and Mogill (1997) developed a framework to describe how young children think in probabilistic situations based on cognitive research that recognizes

developmental stages as well as the recognition of the existence of levels that recycle during stages. The Probabilistic Thinking framework describes children's probabilistic thinking across four levels within four constructs: sample space, probability of an event, probability comparisons, and conditional probability. "Level 1 is associated with *subjective* thinking, Level 2 is seen to be *transitional* between subjective and a naïve quantitative thinking, Level 3 involves the use of *informal quantitative* thinking and Level 4 incorporates *numerical reasoning*" (p. 102). The framework was validated using data obtained from 24 children of grade 1 through 3 and the data was used to refine the framework as well. Their results indicate that students' probabilistic reasoning growth over time, although not following an ordered progression. The framework has been used to develop and evaluate instructional programs in probability.

In developing their framework, Jones et al. (1997) confined themselves to situations that involve theoretical probability (e.g., situations in with symmetry, number, or simple geometrical measures can be used as the basis for determining probabilities). They did not focus on situations where the probability was based of relative frequencies or distributions obtained through experimentation. Their position is that "it is important to begin with children's intuitive thinking in theoretical probability as a basis for subsequent explorations of their thinking in experimental probability" (p. 104). In developing the framework for the current investigation, the author believes it is important to expose students to situations where the theoretical probability is both known and unknown. This will help them develop

notions about sampling and how you can use the data collected through sampling to make prediction and generalizations about underlying populations.

Development of a Probabilistic Inference Framework

The purpose of the framework developed by the author is to characterize the learning of informal statistical inference using a probabilistic context. Although there are several frameworks in the literature, as discussed in the previous sections, there is not a framework to study students' inferential statistical in a probabilistic context. Similar to the frameworks developed by Pfannkuch (2006a, 2006b), Jones et al. (1997) and Vallecillos & Moreno (2002), the framework developed to analyze the data in the current study is based on the SOLO Taxonomy, a general cognitive model developed by Biggs and Collis (1982).

The SOLO (Structure of the Learned Outcomes) Taxonomy postulates that all learning occur in five modes of functioning (Biggs & Collis, 1982):

- *Sensorimotor* (soon after birth) – the individual reacts to the physical environment. This is the mode where young children acquire motor skills.
- *Ikonic* (from 2 years) – the individual internalizes actions in the form of images. This is the mode where children develop words and images that can stand for objects and events.
- *Concrete symbolic* (from 6 to 7 years) – the individual is capable of using or learning to use a symbolic system (i.e. number system) which have an empirical referent. This is the most common mode addressed in learning in the upper

primary and secondary level.

- *Formal* (from 15 to 16 years) – the individual can consider more abstracts concepts and work in terms of "principles" and "theories".
- *Postformal* (possibly around 22 years) – the individual is able to question or challenge the fundamental structure of theories or disciplines.

Most of the students in upper primary and secondary levels are capable of operating within the concrete symbolic mode, although it is possible to have students that still respond within the ikonic mode and other that may respond within the formal mode (Pegg & Devey, 1998).

The progression or growth within each mode is characterized through a learning cycle. The sequence of levels refers to a hierarchical increase in the structural complexity of the responses in a particular mode and can be used to classify the outcomes of learning within any given mode (Biggs & Collis, 1991). The five levels of response within each mode are:

- *Prestructural* (P) – the learner is frequently distracted or mislead by irrelevant aspects of the situation and does not engage the task in the mode involved.
- *Unistructural* (U) – the learner focuses on the problem, but uses only one piece of relevant information.
- *Multistructural* (M) – the learner uses two or more pieces of data without perceiving any relationship between them.

- *Relational* (R) – the learner can now use all data available, integrating each piece of information. The whole has become a coherent structure with no inconsistencies within the known system.
- *Extended abstract* – the learner goes beyond the data and generalizes into new and more abstract features. The learner is reasoning in a new and higher mode of functioning.

Prestructural responses indicate the learner is at a low level of abstraction for the task in question. Unistructural-Multistructural-Relational responses fall within the mode in question. Those levels should be use to describe the point in the learning cycle that the learner has already reached and/or to delineate the desirable learning goals for a particular task (Biggs & Collis, 1991). Extended abstract responses go beyond the level of abstraction expected in the mode of question, which make the response to be extended in to the next level. Table 2.1 illustrates the modes and levels in the SOLO Taxonomy in relation to the concrete symbolic mode.

The three constructs: generalization, variability, and investigation

Comparing the literature on students' informal statistical inference there are similarities as well as differences about the key principles researchers believe are important when investigating how students reason about informal statistical inference. Most of the researchers agree that reasoning informally about statistical inference is being able to make inferences from data without using any statistical formal method or procedure (Pratt, 2000;

Table 2.1. Modes and Level in the SOLO Taxonomy (Biggs & Collis, 1991, p. 65)

Mode	Structural Level	
Formal	5	*Extended Abstract*. The learner now generalizes the structures to tale in new and more abstract features, representing a new and higher mode of operation.
Concrete Symbolic	4	*Relational*. The learner now integrates the parts with each other, so that the whole has a coherent structure and meaning.
	3	*Multistructural*. The learner picks up more and more relevant or correct features, but does not integrate them.
	2	*Unistructural*. The learner focuses on the relevant domains, and picks up one aspect to work with.
Ikonic	1	*Prestructural*. The task is engaged, but the leaner is distracted or mislead by an irrelevant aspect belonging to a previous stage or mode.

Zieffler et al., 2008). Based on our definition of informal statistical inference and important statistical ideas described throughout the chapter, a framework was developed to describe growth in students' abilities to make informal statistical inference in a probabilistic context. Three constructs were identified to be important key principles to be included in the framework: *generalization, variability, and investigation*. These constructs are described in the following paragraphs. Table 2.2 shows the general framework, with a description of what is expected for each level within each construct.

Generalization builds upon the ability students have to move from looking at individual cases to look at the data as an aggregate and beyond. Generalization is making a claim about the aggregate that goes beyond the data (Makar & Rubin, 2009). Statistical generalizations are abstractions from particular cases (sample) that are applied to a broader set of cases (population). Generalizing is making inferences, in contrast with EDA, which

focuses on describing the data at hand. Generalizations are not based on intuitions or previous experience; it has to have data as evidence. This evidence should provide information about whether the claims or predictions they are making are plausible. Students could provide graph, tables, or any other representation they believe would help them support the generalizations they are making. In making generalizations beyond the data students should take into account the size of the samples they are using to draw their conclusion. At higher levels, it is expected that they change their perceptions about the role sample size plays when making inferences as well as how to use the information obtained through the simulations to draw their conclusions and generalizations.

Variability is one of the fundamental concepts in statistics and students need to be aware of it. Inferring from a sample to a population contains elements of uncertainty that can be expressed using the appropriate probabilistic language. Students should suggest the existence of uncertainty by expressing that a prediction is an estimate, or that a conclusion does not apply to all cases (Makar & Rubin, 2009). Students should also be aware that variability exists within a sample as well as between samples. Expressions such as "I got more yellows than greens" or "last time I got more blues, and this time I got more purples" are examples of awareness of variability within and between samples, respectively. It is important that students become aware of both types of variability when making inferences about a population as they increase their level of reasoning.

The *investigation* construct refers to the ability students have to propose an

appropriate investigation process, including describing how the data will be collected and how the results will be analyzed. In a probabilistic context, when students are asked how they will determine the likelihood of an event, they need to think about conducting an experiment to get a sample and use the empirical probability distribution obtained to draw conclusions. In that process they also need to consider the sample size as an important factor. In a probabilistic context students need to be aware of the empirical distribution obtained and its relation to the theoretical distribution. In situations where the data were already collected for the students the investigation process refers to the ability students have to determine how to use the results to draw conclusion about the underlying population.

Research Questions

Most of the probability studies reviewed by Shaughnessy (1992) were conducted with either elementary or college level students. As indicated by Jones et al. (2007) after Shaughnessy's review, the research of probabilistic thinking in the secondary level grew significantly, but researchers focused their studies in students' probabilistic reasoning prior to instruction. As a result, Jones et al. called for the need of research that "investigate the effect of instruction on secondary students' probabilistic thinking" (p. 944). Another important aspect is the use of technology in the teaching and learning of probability and informal inference. Although several researchers have supported the use of technology tools (e.g., Pratt, 2000; Stohl & Tarr, 2002), more research is still needed on how technology influence students' probabilistic conceptions.

Table 2.2. Probabilistic Inference Framework

	Pre-structural	**Uni-structural**	**Multi-structural**	**Relational**
Generalization (G)	• Makes generalizations based on their previous experiences and their intuitions. Ignores sample size.	• Makes generalizations based on a single part of the information given ignoring the others.	• Makes generalizations based on most of the given information or data collected but still ignoring others.	• Make generalizations based on all the information given or the data collected. Recognizes the importance of large samples to make generalizations.
Variability (V)	• Attributes variability to luck or chance in all contexts. No recognition that variability can be described/controlled.	• Recognizes variability among individual results within a single sample (e.g., I got more 5's than 3's). • Has low expectation of variability from an expected distribution based on some theoretical probability distribution.	• Recognizes variability among individual results within a single sample and recognizes variability across different samples (e.g., last time I got more 5's, this time I got more 1's). • Has a better sense of expecting some variability from theoretical probability distributions, but may expect too much variability.	• Recognizes variability within and across samples and relates variability to characteristics of the underlying probability distribution.
Investigation (I)	• Does not propose a data collection process • Believes sampling does not tell you anything about the underlying population.	• Proposes a data collection process, either appropriate or inappropriate (e.g., uses a small sample size), and does not describe how to assess the results.	• Proposes an appropriate or inappropriate data collection process and describes how the results could be analyzed, although not using the most appropriate approach.	• Proposes an appropriate data collection process (e.g., a large sample size) and describes how the results will be analyzed (e.g., from the distribution of the outcomes)

In their review, Jones et al. (2007) generated an agenda for research in probability for the next 10 to 15 years. The necessity of research in order to understand students' conception of theoretical and experimental probability is the first item in this agenda. As indicated by

Jones et al. (2007), although there exist a considerable research in students' conceptions of theoretical probability (e.g., Jones et al., 1999; Watson & Moritz, 2003; Vallecillos & Moreno, 2002) just a few researchers have studied students' conceptions of experimental probability. Even more, studies about students' conceptions about the connection between theoretical and experimental probability are limited (e.g., Aspinwall & Tarr, 2001; Pratt, 2000, 2005; Pratt et al., 2008; Stohl & Tarr, 2002).

Our study includes six case-study students as they are engaged in a series of tasks in a 12-day instructional program in probability. A sequence of tasks focused on similar constructs was analyzed in order to study how students' abilities to engage in probabilistic inference grow. This study intends to bring some insights on how students learn important concepts in informal statistical inference using probabilistic situations, and how that learning is influenced by instruction. The author will pay special attention to students' abilities to make probabilistic inferences taking into account the role of variability and sample size, their ability to make generalizations, and, when possible, their abilities to propose an investigation process. Focused on students' probabilistic inference, the specific questions the author aim to answer are:

1. How do middle school students' abilities to make probabilistic inferences change over the course of instruction?
2. How do middle school students' perception about the importance of sample size when making probabilistic inference change?

CHAPTER 3

METHODOLOGY

A case study methodology (Baxter & Jack, 2008; Stake, 1995; Yin, 2003) was used to conduct this study. A case study allows the researcher to study the complexities of a single case and understand, "its activity within important circumstances" (Stake, 1995, p. xi). This study investigated middle school students as they were engaged solving several tasks in a 12-day instructional program in probability using a computer-based simulator. In this study a case is defined to be a student, specifically six students were selected to be six individual case studies during the investigation. The purpose of this chapter is to describe how the larger study was conducted, the context of the study, the participants, the description of the tasks, the data sources, as well as the methods used to analyze the data.

The Larger Study

As stated by Chance et al. (2007) the use of simulations enhance students' understanding of abstract concepts such as variation and sampling distribution. Students' understanding of such concepts can be developed by carrying out repetitions, controlling parameters (e. g., number of trials, number of samples, weights of events), and explaining the behavior they observe (Chance et al., 2007). Six problem-based tasks were used during the 12 days of instruction where students were required to use a computer-based simulator to collect, display, and analyze data. Using *Probability Explorer* students were also required to

draw inferences and use argumentation and data-based evidence to support their claims. Also, whole class discussions were designed to elicit students' reasoning about the data they empirically collected and the probability distribution they expected based on a sample space.

Probability Explorer was designed as an open-ended learning environment (OELE) with multiple ways to represent data that engage students in designing, simulating, and analyzing results of probability experiments. A well-designed OELE enables learners to build and test their intuitive notions in an exploratory manner. In the software data is represented with randomly generated icons that can be stacked (as a pictograph), lined up in the order they occurred, and listed in a table with the history of all trials. When using *Probability Explorer*, users need to specify the number of trials, which is an important factor for students to consider when making inferences from a sample. In addition, the software enables students to represent the data in multiple ways which help them develop their abilities to choose the best representation tool to analyze the data. Data can be viewed in a pie graph (relative frequency), bar graph (frequency), and data table (frequency and relative frequency). Having multiple ways to represent the data students have the opportunity to observe that some representations differ in the information that is expressed (e.g., frequencies in a bar graph versus the relative frequency table) or show the same information in different ways (e.g., frequencies in iconic pictogram, bar graph, and frequency table).

Possible instructional tasks that utilized *Probability Explorer* as a primary investigation tool were drafted by Hollylynne Lee and James Tarr. These instructional tasks

Table 3.1. Instructional tasks used in larger study (from Stohl & Tarr, 2002, p. 324)

Days of unit	**Task**	**Intent of task**
1-2	Fair Coin Tosses	To discuss the concept of fairness, randomness, and begin to compare empirical results with theoretical probability. To analyze data from a fair coin toss with real coins and tools in *Probability Explorer* such as: stack, lineup, bar graph, and data table.
2-3	Fair Die Tosses	To analyze data from a fair die toss with real die and *Probability Explorer* tools. To discuss outcomes of a die toss for small and large trials. To compare variability of results between a coin toss and die toss to establish the number of outcomes as a factor in "evening out."
4-5	Mystery Marble Bag	Students will analyze data from sampling with replacement from a bag of marbles. Given a bag of 10 marbles with two of six possible colors but unknown distribution of colors, students will use *Probability Explorer* tools to simulate and analyze data and infer bags' contents. Given a bag of 12 marbles with three to six possible colors but unknown distribution of colors, students will use *Probability Explorer* tools to simulate and analyze data and infer bags' contents. To use evidence from simulations to support inferences made about contents of bags of marbles.
6-7	Mystery Fish in a Lake	Given a lake with two different types of fish and an unknown total population, students will use the *Probability Explorer* tools to simulate and analyze data to infer the population ratio of fish and estimate the likelihood of catching each type of fish in a lake. Students will explore several mystery lakes with ratios for three fish and discuss connections between the ratio, probability, pie graph, and percents. (1:2:2, 1:2:3, 2:6:8, and 2:2:6).
8-9	Designing a Model: Weather and Spinner Simulations	Given both a discrete (weather) and a continuous (spinner) probability situation, students will use the Weight Tool to model the situation in *Probability Explorer*. Students will use proportional reasoning to justify why two sets of weights are equivalent, as well as why a given set of weights accurately models a spinner. Students will collect data and analyze the empirical distribution as evidence of whether their Weight Tool is an accurate model.
10-12	Schoolopoly Task	Students will use *Probability Explorer* to simulate rolls of a die and display data using a variety of representations. Student will draw inferences regarding the fairness of a die and estimate theoretical probabilities based on the outcomes of their experiments and simulations. Students will evaluate the validity of arguments and claims based on data.

were created based on results of prior research, a 3-day pilot study, and the researchers' experience with students' understanding and reasoning about probability (Stohl & Tarr, 2002b). After each day's lesson the teachers-researchers revised the learning trajectory they had hypothesized based on students' understanding and perturbations that occurred during that day's lesson. The initially planned tasks were adjusted accordingly. The six problem-based tasks used during the instruction are described in Table 3.1.

Context for the Study

The data reported in this study is part of the larger study that took place in a public middle school in the southern United States. A 12-day instructional program engaged an average-level 6th grade mathematics class (n=23) in a probability unit during the first month of classes. All 23 students in the class were seated in pairs or groups of three at tables with a PC laptop, calculators, and manipulative materials (e.g., dice, spinners). Six students were carefully selected to be used as case studies. These six students were seated in pairs and their computers were connected to a PC-to-TV converter to video-record their computer interactions while microphones captured their conversations. In addition, there was a video camera focused on the three tables to capture students' social interactions with each other and the teachers-researchers.

The probability unit used during the instructional program was designed and taught by Hollylynne S. Lee and James E. Tarr. During instruction, students used real objects and a simulation tool to explore the data. Real objects included coins, die, and spinners and the

simulation tool used was *Probability Explorer* (Stohl, 1999-2005).

During the first two lessons students were given real coins and dice and were asked to conduct experiments in order to determine fairness. During the last part of lesson one and lesson two students had access to the simulation tool *Probability Explorer*. Among the objectives of these two lessons were to analyze data collected by using real objects as well as *Probability Explorer* and to discuss outcomes of a coin/die toss for small and large number of trials. During the instruction special emphasis was given to the importance of large sample size when conducting experiments to make predictions. Instructors also emphasized the use of data as evidence to support their claims. One of the main foci during the instructional unit was the connection between empirical and theoretical probability.

During the six lessons, students were asked to conduct simulations using *Probability Explorer* and to complete worksheets with their pairs. These worksheets included instructions and detailed tables that, in some occasions, included the number of trials to run per experiment. For example, students were asked to run 10 trials six times, then 20 trials six times, then 30 trials six times, and by the end they had a table with six rows but no indications of number of trials to run. After each table, some questions were included that asked students to make predictions based on the data collected.

Participants

Prior to instruction, three pairs of students were selected for the case studies. The selection was based on the scores of two tests the students took: a standardized mathematics

achievement test, and a pretest on probability concepts. The pretest was developed by the researchers and it included items such as determining a priori probabilities and using data to make statements regarding probability. The students were chosen by their relatively consistent ranking on both tests. To do this, the scores were grouped in thirds (high, middle, low) and six students were then chosen to be collectively representative of gender and ethnicity of the class. The six students included four boys (two Caucasians, one Hispanic, and one African-American) and two Caucasian girls. Figure 3.1 shows how the six students were distributed in the high-, average-, and low-scoring groups.

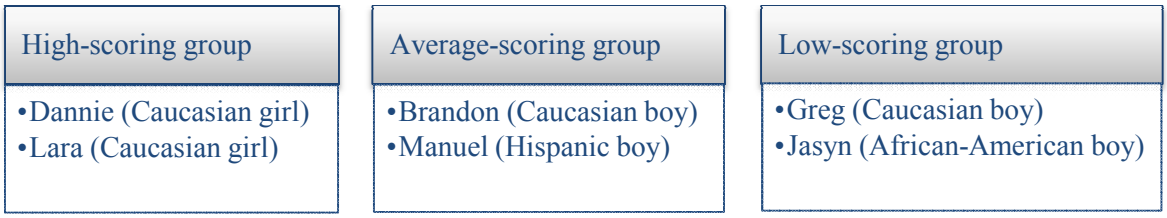

Figure 3.1. Pairs of students selected accordingly with pretests (All names are pseudonyms)

Sources of Data

Six sources of data were collected in the larger study: a pre-interview and a post-interview of the six case study students, all students' worksheets and homework (n=23), a pre-test and a post-test, and a retention test for all students. The order in which the test and the interviews were conducted is shown in Figure 3.2.

The pre-test was used to obtain evidence about students' knowledge of probability and students' informal statistical inference before having the instruction. It consisted of 35 questions, including multiple-choice and open-ended items, covering areas such as fractions

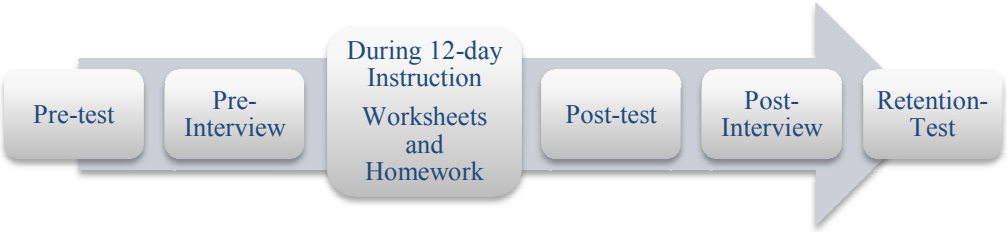

Figure 3.2. Order in which the data were collected

and percents, determining a priori probabilities, proportional reasoning, and using data to make statements regarding probability. This pre-test, in combination with a standardized mathematics test, was used to select the six case-study students prior to the instruction, as described earlier.

After the pre-test, but before the instruction started, semi-structured interviews were conducted with the six students. Semi-structured interviews consist of a set of predetermined questions but the interviewer can be flexible and probe the interviewee based on interactions during the interview (Blee & Taylor, 2002). Semi-structured interviews are useful since they provide greater breadth and depth of information, the opportunity to discover the respondent's experience, and access to people's ideas and thoughts on their own words (Blee and Taylor, 2002, p. 92). The pre-interview consisted of ten tasks about probability and lasted, on average, 40 minutes. The pre-interviews were intended to get a sense of the students reasoning of probability concepts such as probability of an event, sample size, and equally likely events.

During the instruction students worked in their groups and completed a series of

worksheets and homework that were kept as evidence of each student's work. Videotapes of the six case study students' interactions were also collected as well as recordings of their manipulations with the data when using *Probability Explorer*. Videos about students' interaction were not taken into account for this study and only a sample of the worksheets and homework collected were used to investigate students' changes in probabilistic inference.

The post-interview consisted of eleven tasks and lasted an average of 34 minutes. This interview was designed to gather information about the student's thinking when solving probabilistic tasks as well as to investigate whether there were changes in student's informal statistical inference after the instruction in comparison with what the student did at the beginning and during the instruction.

Immediately after the instructional sequence, on day 13, a post-test was given that was parallel in construction to the pre-test but also included several new questions concerning making inferences from data. Ten weeks after the instructional program was over, a retention-test was administered to the 23 students. There were a total of 28 items that matched an item in the pre-test as well as in the post-test. The retention-test was intended to investigate what the students still remembering after the instruction. The following section includes a description of the tasks that were used as well as what was intended by using each task or part of a task.

Description of the Tasks Used to Conduct the Analysis

For this study on students' development of informal statistical inference, the author examined all sources of data from the larger study and selected a total of 13 tasks to use as part of the analysis in this investigation. A visual representation of the number of tasks selected from each source of data is presented in Figure 3.3.

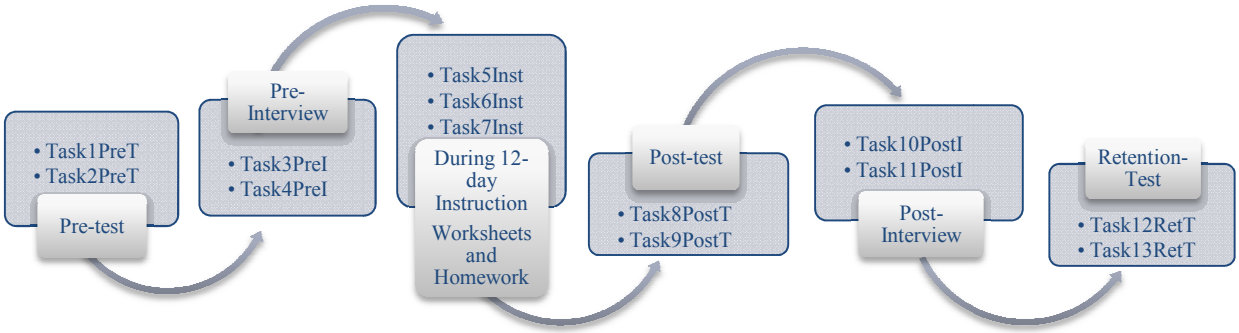

Figure 3.3. Number of tasks chosen from each source of data

As shown in Figure 3.3, two tasks from the pre-test, two tasks from the pre-interview, three tasks from the instructional unit, two from the post-test, two from the post-interview, and two from the retention-test were selected to be analyzed. A description of the concepts/constructs assessed in each task is included in Table 3.2.

What follows is a detailed description about the 13 selected tasks identified in Table 3.2. Four open-ended items were selected from the pre-test and grouped to be considered as two tasks for the analysis (see Appendix A). These items were intended to gather information about students' reasoning of fairness using sample size and the probability distribution

obtained from the sampling. Task1PreT was situated in a context about an old buffalo coin Kiki found in the side of the street. The students were asked if the coin Kiki found was fair or not based in the result of certain number of tosses. Starting with data from 10 tosses, then 100, then 1000 tosses, students were asked to decide if the coin was fair. Task2PreT, included in Appendix A, asked students if they found an old coin and needed to estimate the probability of the coin landing on heads, what they should do. It was expected students proposed to do some data collection and to reason based on the distribution of the results.

Table 3.2. Tasks and items selected from the sources of data

	Data Source of Task	**Task Chosen**	**Context**	**Concepts/Constructs**
Task1PreT	Pre-Test	Item 31-33	Old buffalo nickel found on the street. Is it fair?	• Importance of sample size in order to make generalizations beyond the data collected. • Fairness • Variability within and across samples.
Task2PreT	Pre-Test	Item 34		• Student should propose a data collection process (sample size larger than 50) and state how he would analyze the results (looking at the distribution).
Task3PreI	Pre-Interview	Task 4 – Mystery Bags of Marbles	Given the total number of marbles, predict the proportion of each color.	• Given the total number in the bag, predict its content. • Importance of sample size in order to make generalizations beyond the data collected.
Task4PreI	Pre-Interview	Task 5 – The Plastic Cup	Estimate the likelihood a plastic cup would land upside down.	• Estimate proportion of unknown population. • Importance of sample size in order to make generalizations beyond the data collected.
Task5Inst	Day 4 – Lesson 3: *Mystery Bag of Marbles*	Worksheet – Part I-V	Bag of Marbles with known total, but unknown content	• Given the total number in the bag, predict its content using PE • Importance of sample size in order to make generalizations beyond the data collected.

Table 3.2. Continued

Task6Inst	Day 6 – Lesson 4: *Mystery Fish in Lake*	Worksheet – Part I-III	Fish in Lake with 2 types of fish, but unknown total	• Estimate proportion of unknown population. • Importance of sample size in order to make generalizations beyond the data collected.
Task7Inst	Day 10 – Lesson 6: *Schoolopoly*	Poster	Decide whether or not the dice produced by certain companies were fair or not	• Estimate proportion of unknown population • Importance of sample size in order to make generalizations beyond the data collected. • Variability within and across samples
Task8PostT	Post-Test	Item 20-22	Bottle cap found in the sidewalk. Is it fair?	• Importance of sample size in order to make generalizations beyond the data collected. • Fairness • Variability within and across samples.
Task9PostT	Post-Test	Item 23		• Student should propose a data collection process (sample size larger than 50) and state how he would analyze the results (looking at the distribution).
Task10PostI	Post-Interview	Task 4 – Mystery Jars of Jolly Ranchers	Two flavors in the jar. Proportion of each flavor?	• Given the total number in the bag, predict its content using PE • Importance of sample size in order to make generalizations beyond the data collected
Task11PostI	Post-Interview	Task 5 – The Toothpaste Cap	Drop the toothpaste cap to estimate likelihood the cap would land upside down	• Estimate proportion of unknown population. • Importance of sample size in order to make generalizations beyond the data collected.
Task12RetT	Retention Test	Items 22-24	Unusually shaped button, flat on one side and with ridges in the other side. Is it fair?	• Importance of sample size in order to make generalizations beyond the data collected. • Fairness • Variability within and across samples.
Task13RetT	Retention Test	Item 7		• Student should propose a data collection process (sample size larger than 50) and state how he would analyze the results (looking at the distribution).

In the pre-interview only two tasks were selected to be used to gather information about students' understanding of inference. These tasks and the protocol followed by the interviewers are included in Appendix A. Mystery Bags of Marbles (Task3PreI) consisted of two bags, Bag A contained 10 marbles with only two colors present (blue and red), and Bag B contained 10 marbles with four colors present (blue, red, green, and yellow). Students were asked how many times they would need to draw a single marble with replacement in order to be very confident about how many marbles of each color are in the each bag. This task was intended to have students think about sample size and its importance to be able to make generalization beyond the data collected. The second task chosen, The Plastic Cup (Task4PreI), uses a non-deterministic situation where the students needed to determine the number of times they would need to drop a plastic cup in order to be confident of their estimate about the likelihood that the cup landed upside down. This task intended to address the importance of sample size in order to make generalizations beyond the data collected.

Task5Inst (Mystery Marble Bag) provided students a situation where the total number of marbles in the bag (N) was given, but no information about the content was stated. In this task students were asked to collect data and draw inferences about the theoretical distribution of the marbles in the bag. The task started by indicating to students the number of trials to run, but at the end students had to determine the number of trials to run. The total number of marbles in the bag was purposely set to ten to promote students uses of percentage and proportional reasoning. The task is included in Appendix A. Part I asked students to run ten trials six times, in part II they needed to run 20 trials six times and in part III they had to

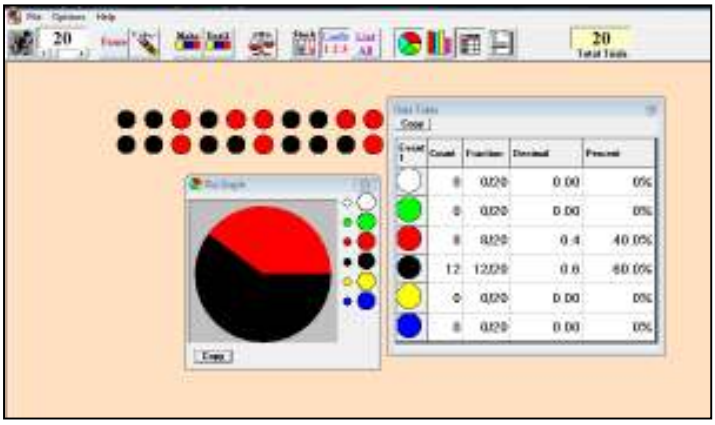

Figure 3.4. Use of a pie graph and a data table to make inferences about of the content of the bag of marbles after 20 trials

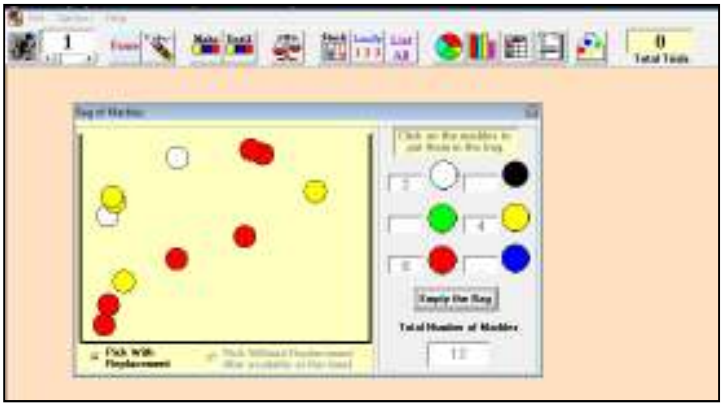

Figure 3.5. Designing a bag of marbles

choose the number of trials to run (see Figure 3.4). Part IV asked the pair of students to work as a team. One of them had to design the content of his bag while the other tried to "Guess my bag" (see Figure 3.5). In all parts they were asked to predict the content of the bag and explain how confident they were about their prediction. For the purpose of the analysis of this investigation only data from Part IV was used, data from Part I through III was used if completely answered by the students. During this task students should be able to notice the

variability among samples obtained from the same data. They were asked to arrive to conclusions about the content of the bag and to provide data-based evidence to support their conclusions.

Task6Inst (Mystery Fish in a Lake) was designed to induce a perturbation about how to determine the number of trials to run and to know when enough data had been collected. The task is given in Appendix A including parts of the worksheet students completed. The total number of fishes in the lake was unknown, but they did know there were two types of fish and they were asked to estimate the probability of catching a particular type of fish. The task also encouraged students' social negotiations about whether they were confident that enough data had been collected to estimate the relative distribution of fish and the probability of catching a particular type of fish. Parts I, II, and III were used as parts as the analysis and grouped to be considered one task. Similar to the parts selected from Task5Inst, students were asked to run ten trials six times, then 20 trials six times, and at the end students choose how many trials and how many times they wanted to run the simulation. Students were asked to estimate the probability to catch a Blue Bass fish based on the data they collected (see Figure 3.6).

Task7Intst (The Schoolopoly Task) was the final task of the unit. During this task students had to collect evidence to infer whether or not a dice company produces dices with equiprobable outcomes (see Appendix A). In this task, students were asked to consider a rumor that some of the die may be biased, to justify if they found enough evidence to retain

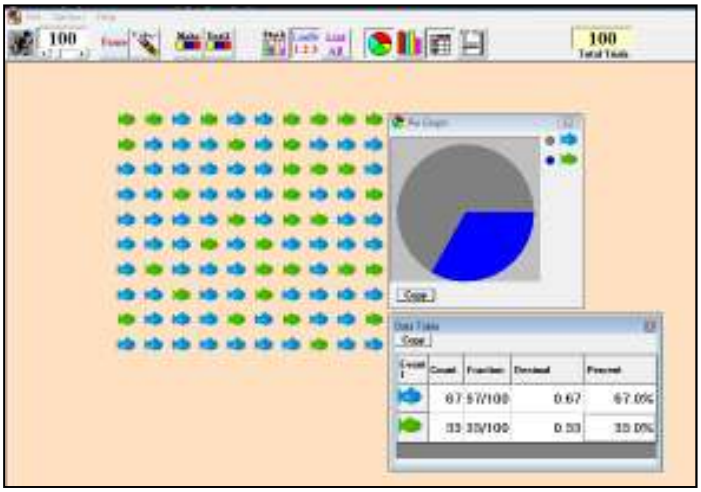

Figure 3.6. Mystery Fish in a Lake outcomes after 100 trials

an equiprobable hypothesis or to claim that the die had unequiprobable outcomes, and propose an estimate for the probability distribution (see Figure 2.2). The students had to decide how much data to collect, how to arrange the iconic form of the outcomes on the screen, how to display data numerically and graphically as they analyzed data, and what data sets and representations to use to make and support claims about the whether the die was fair with equiprobable outcomes.

During this final task students used *Probability Explorer* to simulate rolls of a die and display data using a variety of representations. They needed to draw inferences regarding the fairness of a die and estimate theoretical probabilities based on the outcomes of their experiments and simulations. The weight and theoretical probabilities for events 1-6 in each of the three companies the three pairs worked on are presented in Table 3.3. The three pairs worked on dice from Dice R' Us, High Rollers, Inc., and Slice n' Dice.

Table 3.3. Weights and theoretical probabilities (Lee, Angotti & Tarr, 2010, p. 76)

Company Name	**Weight**	**Weight**	**Weight**	**Weight**	**Weight**	**Weight**
	[P(1)]	***[P(2)]***	***[P(3)]***	***[P(4)]***	***[P(5)]***	***[P(6)]***
Dice R' Us	2	3	3	3	3	2
	[0.125]	*[0.1875]*	*[0.1875]*	*[0.1875]*	*[0.1875]*	*[0.125]*
High Rollers, Inc.	2	3	2	3	2	3
	[0.133]	*[0.2]*	*[0.133]*	*[0.2]*	*[0.133]*	*[0.2]*
Slice n' Dice	4	5	5	5	1	5
	[0.16]	*[0.2]*	*[0.2]*	*[0.2]*	*[0.04]*	*[0.2]*

For this task students had to create a poster with the evidence they found to support their claim. A screenshot of their posters is presented in Figures 3.7-3.9. Since students did not complete a worksheet during this task the poster will be used to analyze their probabilistic inference.

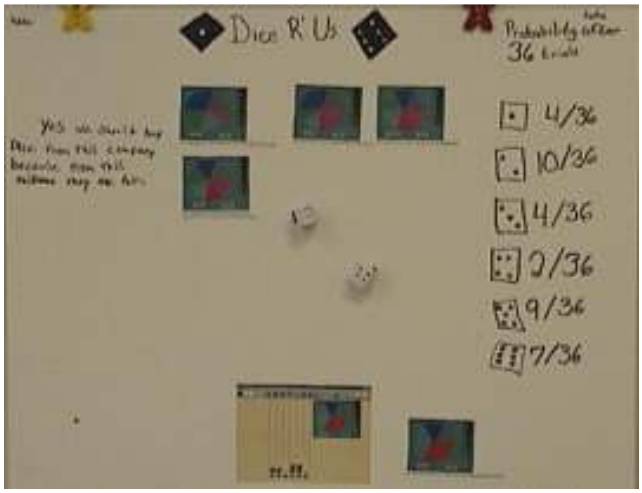

Figure 3.7. Dannie's and Lara's Poster - Dice R' Us

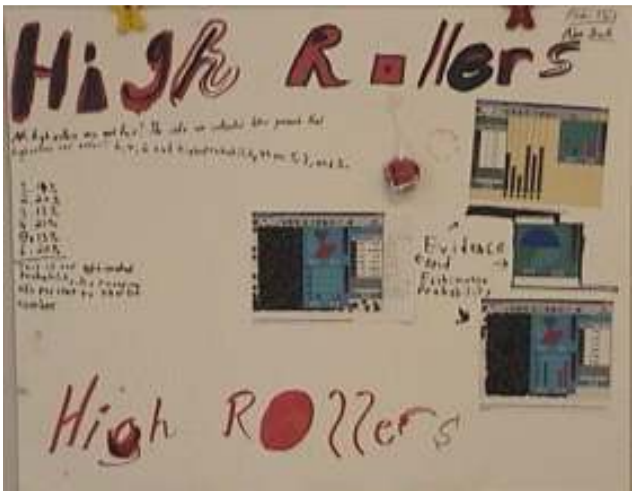

Figure 3.8. Manuel's and Brandon's Poster - High Rollers

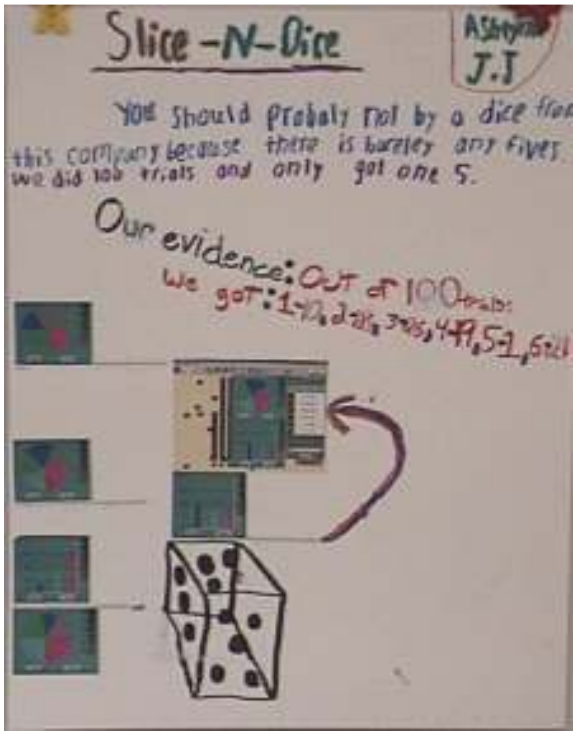

Figure 3.9. Greg's and Jasyn's Poster - Slice -N- Dice

Four open-ended items were selected from the post-test and similarly grouped into

two tasks (see Appendix A). These items were parallel to the four items selected from the pre-test. In Task8PostT the students were told that Alicia found a bottle cap in the sidewalk and that she will use it to decide if she or her brother will use the computer first when they get home. Students were asked whether the bottle cap was fair or not. Results were given for tossing the bottle cap 10 times, 100 times, and 1000 times. The students were expected to state that the bottle cap was unfair based on the distribution or that with 100 and 1000 times you should see a trend toward outcomes being close to even. Task9PostT asked students how many times they need to toss a bottle cap in order to determine if it is fair or not. Students should propose they will need to collect some data and specify that more than 50 trials would be needed. They were also expected to reason based on the distribution of the results.

Two tasks were selected from the post-interview (see Appendix A). These tasks are similar in construction with the two tasks selected from the pre-interview. Task10PostI (Mystery Jars of Jolly Ranchers) and Task11PostI (The Toothpaste Cap) had students think about sample size, similar to Mystery Bags of Marbles (Task3PreI) and The Plastic Cup (Task4PreI) tasks in the pre-interview. In the Mystery Jars of Jolly Ranchers, the students were asked how many times they would need to draw a single piece of candy, replacing it after each draw, to be very confident about how many candies of each flavor were in Jar A and in Jar B. They only knew each Jar had 10 Jolly Ranchers and that Jar A had two flavors while Jar B had four. The Toothpaste Cap task presented a non-deterministic situation where the students were asked to estimate the likelihood that the cap landed upside down. Students needed to determine the number of times they would need to drop the cap in order to be

confident of their estimate.

Similar to the pre-test and the post-test, four open-ended items were selected from the retention-test and grouped into two tasks (see Appendix A). In Task12RetT the students were told that Nathalie found an unusually shaped bottom, flat one side and with several ridges in the other side, while playing at the park. She told her brother that if she tossed the button she thought it was more likely to land with the ridges side up. Results were given after tossing the button 10 times, 100 times, and 1000 times. The students were expected to reason about the fairness of the button based on the distribution and be able to state that the button was unfair because with 100 and/or 1000 times you should see a trend toward outcomes being close to even. Task13RetT asked students how many times they will need to toss an unusually shaped button that was flat in one side and had ridges on the other side in order to determine if it is fair or not. Students should propose they will need to collect some data and specify that more than 50 trials would be needed. They were also expected to reason based on the distribution of the results.

As mentioned earlier, there were a total of six tasks used during the instruction. Each task, divided in different parts, consisted of worksheets that students completed in their groups in the classroom and homework each student completed each day. These six tasks were briefly described in Table 3.1, but only Lesson 3, Lesson 4 and Lesson 6 were considered to be used in this investigation, in part because of lack of responses in Lesson 1 and Lesson 2. The tasks used for the analysis of the data are included in Appendix A.

Methods of Analysis

The data was analyzed to describe how the students' engaged in probabilistic inference at various points before, during, and after instruction. The first step in the analysis process was to search for tasks were students were asked to make statistical inferences. The author searched for similar tasks through all sources of data so that they could be clustered in different groups. The chosen tasks have been described and appear in Appendix A. The tests, as well as the interviews, were parallel in construction, meaning that most items in the pre-test had a parallel item in the post-test as well as in the retention test, and every question on the pre-interview had a parallel question in the post-interview, as well as parallel tasks during the instruction. The parallelism among the instruments helped in being able to see and describe students' changes in their approach to informal inference. The second step was the creation of four sequences based on the constructs each task intended to evaluate (see Figure 3.10). Tables C.1 through C.4, in Appendix C, include the tasks that compound each sequence.

Once the sequences were created, the next step was to develop a theoretical framework. Key elements acknowledged to be essential in informal statistical inference, identified through the review of the literature, were used to characterize each level in the SOLO Taxonomy. Four constructs were included in the initial theoretical framework: generalization (G), variability (V), investigation (I), and sample size (SS).

Once the sequences of tasks were selected and the framework was created, the

investigator started coding the interviews paying special attention to the constructs assessed in the framework. Initially the interviews were recorded in a VHS format, so they needed to be transferred from a VHS to a digital format. Most interviews had been previously transcribed, but the researcher needed to transcribe several of the students' responses to particular interview tasks used in the sequences. As discussed in Chapter 2, the framework was created using the different levels of the SOLO Taxonomy (Biggs & Collis, 1982).

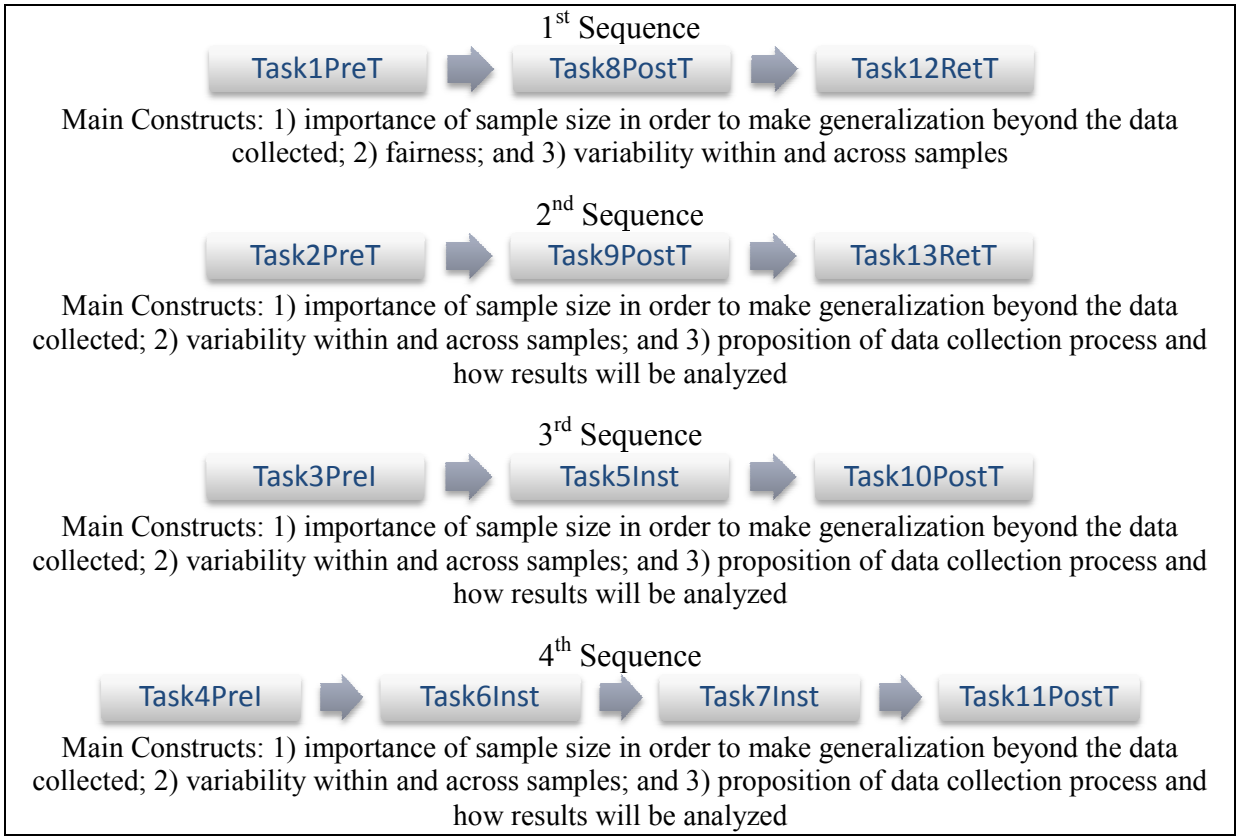

Figure 3.10. Sequences and main constructs used to characterize students' levels

The next step was to validate the theoretical framework. Using the data from one student, the author validated the framework finding the constructs were overlapping one with

the others, specially the sample size construct. This made it hard to use the framework to characterize the level of that particular student in each of the constructs. The process of validating the framework showed inconsistencies about the student level of thinking and the author, in conjunction with an experienced researcher, decided to eliminate the sample size construct and instead use the awareness of sample size as a characteristic within each of the other three construct. The framework was refined using the feedback obtained from the validation process and the final product, the general framework, is presented in Appendix B, Table B.1. The refined framework characterized students levels of reasoning using three constructs: generalization (G), variability (V), and investigation (I). For each sequence only the appropriate constructs were evaluated, meaning that not all tasks were intended to have students dealing with all three constructs (see Figure 3.10). The 1st sequence is the only one in which the investigation construct could not be characterized.

Once the general framework was developed (see Appendix B, Table B.1) individual frameworks were created for each sequence. These tables are presented in Appendix B, tables B.2 through B.5. The individual frameworks were used to characterize students' level in each specific sequence. In these individual frameworks each level and construct were described using the context of the tasks. While the author analyzed the data and characterized students' levels, the framework was slightly enhanced in order to reflect the data in hand. Once the individual frameworks were created, the tasks were evaluated and the students reasoning within each task was assigned to a SOLO Level.

In order to analyze the changes in students' probabilistic inference each sequence was analyzed one at a time. The first and the second sequence gave the author the opportunity to analyze changes in each student individually since the tasks came from the pre-test, post-test, and the retention test. The third and the fourth sequence contain data from individual sources such as an interview as well as data collected during instruction that represent negotiated responses among the partners.

The researcher used peer checking, with an experienced researcher, discussing some of the student's responses and the level assigned, in order to gain some level of reliability. In case of questionable responses, the researcher decided to assign the students to the lower level in question (e.g., if the response where questionable on whether it will be a uni-structural or a multi-structural level, the author located the student at the uni-structural level).

As mentioned previously the data was obtained from individual sources as well as instructional sources where negotiation among peers might occur. For the analysis of those sequences that included instructional sources of written data, the author is assuming there was negotiation among the peers so each student in the pair will be assigned the same level of reasoning during these tasks. Those tasks are Task6Inst and Task7Inst. This is a reasonable assumption given that it was expected by the teacher for students to reach agreement in their work. In addition, a review of the videos of the instructional tasks demonstrates students' active engagement with each other and the software and often taking turns writing responses on the written task sheet.

Once the author and the experienced researcher agreed in the assignment of all students to their respective SOLO levels, a qualitative analysis was conducted within each Case-Study across the sequences. Next, a cross-analysis was conducted to investigate across students' changes within each construct, as well as across the tasks in the order they were collected.

The following three chapters describe findings about the six case-study student's changes in probabilistic inference. Chapter 4 describes the findings about Lara & Dannie. Lara and Dannie were classified as the high-scoring group at the beginning of the instruction. A description of the findings about Manuel and Brandon's probabilistic inference is presented in Chapter 5 and Greg and Jasyn's findings are described in Chapter 6. Finally, a cross-analysis of the cases will be discussed in Chapter 7.

CHAPTER 4

RESULTS: *Lara & Dannie*

As mentioned earlier, there were four sequences used to analyze students' changes in probabilistic inference. Within this chapter there are two sections, each describing the reasoning of an individual. Each section is divided into four sub-sections which describes a student's performance within each sequence. This chapter describes Lara's probabilistic inference changes in the first section and Dannie's probabilistic inferences changes in the second section. Recall that Lara and Dannie are the high scoring group. Tables of the frameworks are included in Appendix B, tables B.1 through B.5 and the tables of the sequences could be found in Appendix C, tables C.1 through C.4.

The case of Lara

Lara's 1st sequence

The first sequence is included in Appendix C, Table C.1. The sequence includes three tasks where students have to reason about the fairness of a coin, a bottle cap, and a shaped button, respectively. The tasks are included in Appendix A. Only two of the three constructs were used to analyze changes in probabilistic reasoning in this sequence: generalization and variability.

Generalization. Lara started reasoning at a pre-structural level when making generalizations about the fairness of a coin. Her answer to Task1PreT could be interpreted as making generalization based on her previous experiences with coins. She believes a coin is

fair even when a big difference exists between the number of heads and the number of tails. As she stated even when a coin had been tossed 1000 times with 643 tails and 357 heads, “it can always be a fair coin unless it always lands on one side”. Lara’s reasoning is not based in sample size at the beginning of the sequence, but as she moved on to Task8PostT she recognized sample size as an important factor to determine the fairness of the bottle cap. After 10 tosses she said “maybe, they only tried it 10 times” recognizing the sample size was small to be used to make an inference about the fairness of the bottle cap. She still was not convinced after 100 tosses and after 1000 tosses she recognized the difference between the times the bottle cap landed up and down was big enough to determine the bottle cap was unfair. At this point during the instruction she recognized sample size is an important factor when making inferences, reasoning at a multi-structural level when making generalizations about the fairness of the bottle cap. At the time Lara took the retention test, Task12RetT, she kept paying attention to the sample size and to the difference among the times the button landed with the ridged side up and down, reasoning at a multi-structural level. After 10 tosses she wrote that the button had not been flipped 50 times yet so she cannot determine if it was fair or unfair. After 100 tosses Lara decided it seems to be unfair because “a difference of 30” was a big difference and after 1000 tosses she definitively determined “yes it is unfair big difference”.

Variability. Lara did not recognize the existence of variability at the beginning of the sequence (Task1PreT) as she believed all coins are fair, reasoning at a pre-structural level. In the second task of the sequence (Task8PostT) she demonstrated recognizing variability as she

is paying attention to the differences between the times the bottle cap landed up and down. There is no evidence she recognized variability between the three samples, but she did recognize the variability among landing up and landing down in a single sample, locating her in a uni-structural level of reasoning in the variability construct. By the time Lara took the retention test she still recognized the variability among the times the button landed with the ridged side up and the times it landed side down, as she did during Task8PostT, keeping her level of reasoning at a uni-structural level within the variability construct.

Table 4.1. Lara's SOLO level in each task – 1st sequence

	Task1PreT	**Task8PostT**	**Task12RetT**
Generalization	Pre-structural	Multi-structural	Multi-structural
Variability	Pre-structural	Uni-structural	Uni-structural

Lara's 2nd sequence

The second sequence is included in Appendix C, Table C.2. This sequence asked to describe what she will do to determine the fairness of a coin (Task2PreT), a button (Task9PostT), and a bottle cap (Task13RetT), and what would give her convincing evidence.

Generalization. Results from the first sequence showed Lara used her previous experiences to generalize that all coins are fair. Since she believes all coins are fair, she expects her coin to be fair as well and, in her reasoning, the only chance a coin has to be unfair is if the outcomes are either all heads or all tails. Her reasoning is at a pre-structural

level since she is using her previous experiences to make a generalization. At Task9PostT she proposed to flip the bottle cap "100 to 1000 times if it is not about the same at 100." Lara is reasoning at a multi-structural level since she is proposing the use of a large sample to make generalizations. Her answer to Task13RetT indicates she will flip the button and make generalizations based on the outcomes, but she is expecting to get half outcomes with the ridged side on top and the other half with the ridged side down. She did not specify how many times she will flip the button, but she stated she will have convincing evidence "if [she] flipped it with the ridge side on top 50 times and [she] flipped it on the other side 50 times", which locate her at a uni-structural level.

Variability. At first Lara does not seemed to be aware of the variability that could happen among the individual results within a single sample. She does not allow variability to occur, expecting the coin to be fair even if there is a big difference between the number of heads and tails in Task2PreT. At this point of the instruction she is reasoning at a pre-structural level within the variability construct. In Task9RetT Lara shows awareness of variability within a sample since she is expecting to get "about the same" after flipping the bottle cap 100 times. This locates her reasoning at a uni-structural level. At the end of the sequence, Task13RetT, she expects to obtain "the ridge side up 50 times and… the other side 50 times" in order for the button to be fair. Here she shows low tolerance of variability in a 50-50 chance situation, locating her at a uni-structural level.

Investigation. Lara proposed a data collection process, although not an appropriate

one at the beginning, and described how she will assess the results in Task2PreT and Task9Post, reasoning at a multi-structural level. At first she proposed to collect a sample of size 10, but by the time she answered Task9PostT she increased the sample size to 100. Moreover she mentioned to flip the coin "1000 times if it is not about the same at 100" to be convinced. Although Lara described what would give her convincing evidence about the fairness of the coin and the bottle cap, she did not describe what would give her convincing evidence about the button, locating her at a uni-structural level in Task13RetT.

Table 4.2. Lara's SOLO level in each task – 2nd sequence

	Task2PreT	**Task9PostT**	**Task13RetT**
Generalization	Pre-structural	Multi-structural	Uni-structural
Variability	Pre-structural	Uni-structural	Uni-structural
Investigation	Multi-structural	Multi-structural	Uni-structural

Lara's 3rd sequence

The third sequence is included in Table C.3. In this sequence Task3PreI asked students to determine the number of times they needed to draw a single marble, out of a bag containing 10 marbles, in order to be very confident about how many marbles of each color are in the bag. The task started with a bag containing 10 marbles and only two colors are present (blue and red), and then was changed to a bag containing 10 marbles but four colors (blue, red, green, and yellow). Task 10PostI was similar to Task3PreI but in a different context, a jar of Jolly Ranchers. Task 7Inst asked to determine the content of a bag of

marbles just knowing that it contains 12 marbles. No information about the colors was given.

Generalization. In the first task, Task3PreI, Lara does not make generalizations about the content of the bag of marbles stating that "[she] would never be sure unless [she] looked" inside the bag, locating her at a pre-structural level in the generalization construct. In Task5Inst she based her generalization on several samples of size 12, where 12 is the total content of the bag. Although Lara used several samples to make generalizations, she used the "total weight approach" (Lee, 2005) to decide on the size of these samples. In addition, although she used the distribution of the outcomes she did not pay attention, for example, to the number of white marbles in all seven samples to decide how many white marbles were in the bag. Instead she used "the biggest number out of the 3" and "subtracted the lower numbers from the biggest" (e.g., from a distribution of 4 whites, 5 reds, and 3 blues she subtracted 5 minus 4 and 5 minus 3). Her final prediction of 1 white marble, 10 red marbles, and 1 blue marble was far away from the underlying population of 4 white, 5 red, and 3 blue marbles. Her reasoning locates her at a uni-structural level since she did not consider sample size and she used the distribution on an inappropriate way. At Task10PostI her level of reasoning is at a multi-structural level. In the first part of the task she proposed to draw a Jolly Ranger "maybe 100 times" because "it is like having 10 but add zeros" and in the second part of the task she proposed to draw Jolly Ranchers 400 times because there are four flavors. Although she used the distribution of the outcomes to determine how many Jolly Ranchers of each color were in the jar, she based her predictions on a single sample. From a

distribution of 40 cherries, 120 grapes, 40 lemons, and 200 green apples she said:

> "without even trying to figure it out I'll know there are probably more green apple than anything else, and then I'm going to say that maybe grapes may comes second, and then there are probably about the same amount of cherries and lemons."

Variability. In Task3PreI she did not seem to recognize variability, reasoning at a pre-structural level in the variability construct. During instruction, Task5Inst, Lara recognized variability within samples noticing the biggest and the smallest number out of the three (white, red, and blue marbles). She did not seem to recognize the variability across the different samples, which makes her reasoning at a uni-structural level. She was also reasoning at a uni-structural level in Task10PostT since she recognized variability among the individual results within a single sample. After the 400 trials she recognized there were probably more green apples than anything else, that probably grapes may comes second, and there were about the same amount of cherries and lemons.

Investigation. Lara started reasoning at a pre-structural level in the investigation construct since she does not propose a data collection process stating that "[she] would never be sure [about the content of the bag] unless [she] looked because you could be pulling out the same one" in Task3PreI. She said "any number wouldn't be good enough" indicating she did not recognizes sampling as a tool to estimate the content of the jar of Jolly Ranchers. Her answer was consistent during the second part of the task when there were four colors, instead of two, present in the bag. In Task5Inst she ran seven simulations of size 12 each. Here she

was using the "total weight approach" (Lee, 2005) since she knew the bag has 12 marbles in total. Also, her predictions about what was in the bag were not based on proportions or percentages, but using the outcomes of a single sample and subtract the biggest outcome from the smallest outcomes. Her level of reasoning is at a multi-structural level. In Task10PostI she proposed to draw a single marble 100 times for the first part of the task (when there were two colors) and 400 times for the second part (when there were four colors). She reasoned appropriately on how to assess the results using the distribution of the outcomes to make generalizations about the jar of Jolly Ranchers after drawing a single candy 400 times (for details refer to the generalization construct analysis above). At his point she was reasoning at a relational level since she proposed to use a large samples size and assessed the results using the distribution of the outcomes.

Table 4.3. Lara's SOLO level in each task – 3rd sequence

	Task3PreI	**Task5Inst**	**Task10PostI**
Generalization	Pre-structural	Uni-structural	Multi-structural
Variability	Pre-structural	Uni-structural	Uni-structural
Investigation	Pre-structural	Multi-structural	Relational

Lara's 4th sequence

The fourth sequence is included in Appendix C, Table C.4. This sequence is the longest including four tasks. Task4PreI had students reasoning about the likelihood of a plastic cup to land upside down. Students were asked to determine how many times they

needed to drop the cup in order to be very confident in their estimates. Task6Inst presented a situation about a lake that has two types of fish in it, but no one knows what the proportions of each fish are. They needed to estimate the probability to catch a Blue Bass. In Task7Inst students needed to determine whether or not the dice sold by a given company was fair or not. Students presented their results in a poster session. Task11PostI was similar to Task4PreI but using a toothpaste cap instead of a plastic cup.

Generalization. In Task4PreI, Lara started reasoning from her previous experiences with plastic cups stating it is "not very likely" for the plastic cup to land upside down "because there have been times [she has] dropped all the cups [and] usually they've landed on their side." She also mentioned that "it is probably just as likely to land [upside down] as like [right side up]" stating also that "maybe out of 10 tries it might possibly [land upside down] once… if you ask me out of 50 tries, it might do it once." When re-questioned she used a sample of size 100 to make generalizations about the likelihood of the plastic cup to land upside down, although she never mentioned how she will make sense of those 100 trials. Since she tried to make generalizations using samples of size 10 and 50 and she proposed to use a sample of size 100 when re-questioned about how many times she will need to drop the plastic cup to be pretty confident in her estimate, her reasoning is at a uni-structural level. During Task6Inst, Lara had to determine the probability that a fish caught in a lake was a Blue Bass. She started by collecting some samples of size 10, but makes no generalizations based on those samples, but then she makes generalizations about the probability based on

several samples of size 20. Then she used the results from those eight samples of size 20 and "took the averages and turned it into a fraction." Her estimation that "3 out of 10 catches" would be a Blue Bass was pretty close to the theoretical probability of 1 out of 3 catches. Her level of reasoning increased to a multi-structural level.

During the Schoolopoly task, Task7Inst, Lara investigated the fairness of the dice sold by Dice R' Us with her partner Dannie. Lara (same as Dannie) recommended the school "should buy dice from this company because from this evidence they are fair." She used several samples of different sizes, ranged from n=6 to n=100, and used the distribution of the outcomes to make generalizations. Her level of reasoning is at a multi-structural level, since she is not making the connection between sample size and generalizations. From the smallest sample (n=6) she generalized the dice "doesn't look fair" but after 36 and 100 trials she said "it was pretty even" so she thinks the dice was fair. She used the empirical probability distribution after 36 trials as her estimation of the theoretical probability of each outcome. The dice sold by Dice R' Us were after all unfair, but with a pretty even distribution (see Table 3.3).

After instruction, in Task11PostT, Lara dropped the toothpaste cap 300 times in order to be confident in her estimate of the likelihood that it will land upside down. She also mentioned the distribution of the outcomes in a specific case stating that "there is three ways it could landed, so [she] will say 300 times because if it where even it will come out a 100, a 100, a 100." She also mentioned that 300 were "a lot of tries" and that the outcomes should

be a representation of the underlying population stating "it should probably come out what it should." Lara's level of reasoning in the generalization construct is at a multi-structural level since she based her generalizations in a large sample and, although proposed using the distribution of the outcomes to make generalizations, she based her estimation in the counts instead of observing the percentages. She also seemed to recognize the importance of large samples to make generalizations.

Variability. Lara did not seem to recognize variability in Task4PreI. She seems to be expecting the same results from a sample of size 10 and a sample of size 50 (1 out of 10 and 1 out of 50), not recognizing that variability could be affected by sample size. In this task her level of reasoning is at a pre-structural level. In Task6Inst Lara (same as Dannie) recognized the variability across different samples as well as variability within a single sample. Her level of reasoning is at a multi-structural level. As she recognized the variability across samples, she "took the average and turned it into a fraction" in order to determine the likelihood to catch a Blue Bass fish and she observed there were more Green Gills than Blue Bass in each single sample. In Task7Inst, Lara recognized the variability within a single sample as she noticed that after six trials there were no 3's and no 6's in the outcomes. After 36 trials she noticed "it was pretty even" and mentioned the dice were fair with a certain amount of trials, noticing the difference on the outcomes from different samples. She also expected too much variability stating the dice were fair even after getting 4 ones, 10 twos, 4 threes, 2 fours, 9 fives, and 7 sixes out of 36 trials. Her level of reasoning here is at a multi-structural level. In Task11PostI she is proposing to drop the toothpaste cap 300 "because if it is even it will

come out 100, 100, 100" indicating she is not leaving space for variability in equiprobable distributions in order to claim something is fair. Lara's reasoning at this point is uni-structural.

Table 4.4. Lara's SOLO level in each task – 4th sequence

	Task4PreI	**Task6Inst**	**Task7Inst**	**Task11PostI**
Generalization	Uni-structural	Multi-structural	Multi-structural	multi-structural
Variability	Pre-structural	Multi-structural	Multi- structural	Uni-structural
Investigation	Uni-structural	Multi- structural	Multi- structural	Multi- structural

Investigation. In the investigation construct Lara proposed the collection of 100 trials but did not mention how she will use those 100 trials to assess the results in Task4PreI. Her level of reasoning here is uni-structural. During instruction, Task6Inst, Lara and her partner collected several samples and used the distribution of the outcomes, specifically "[they] took the average" of the outcomes, to assess the results. Since the samples proposed were of small size, she is reasoning at a multi-structural level. In Task7Inst, the Schoolopoly task, Lara collected several samples of different sizes and used the distributions of the outcomes to conjecture the fairness of the dice. This reasoning situates her in a multi-structural level. In the last task of this sequence, Task11PostI, Lara proposed to collect a sample of an appropriate size (n=300) and make implicit reference to the use of the distribution of the outcome. She said "[she] will say 300 because if it were even it will come out a 100, a 100, a 100 and that way [she] will be able to tell." She is reasoning at a multi-structural level in the

investigation construct.

Lara started reasoning mostly at a pre-structural level during the *pre* instructional tasks, with the exception of Task2PreT and Task4PreI in the investigation construct. During instruction Lara demonstrated to be reasoning mostly at multi-structural level, except Task5Inst. After instruction her level of reasoning was steady at a multi-structural level in the generalization construct whereas she showed to be reasoning, mostly, at a uni-structural level in the variability and investigation constructs. In summary, Lara demonstrated to increase her levels of reasoning to a multi-structural level in all four sequences in the generalization and the variability constructs. In the investigation construct, her level of reasoning decreased to a uni-structural level in the 2nd sequence, and increased to a relational level in the 3rd sequence, remaining steady at a multi-structural level in the 4th sequence.

The case of Dannie

Dannie's 1st sequence

The first sequence is included in Appendix C, Table C.1. This sequence includes three tasks where students have to reason about the fairness of a coin, a bottle cap, and a shaped button, respectively. Only two of the three constructs were used to analyze changes in probabilistic reasoning in this sequence: generalization and variability.

Generalization. Dannie started reasoning at a uni-structural level in the generalization construct, recognizing that the coin was unfair because the outcomes "should be closer to even", but she did not pay attention to the sample size. She did not seem to recognize that

small samples are not appropriate to make inferences. Dannie's reasoning during Task8PostT improved to the next level. She recognized that "it's a small amount of times you flipped it" when the bottle cap was tosses only 10 times, and after 100 and 1000 tosses she mentioned that the bottle cap was unfair because the numbers "should have been closer to be fair." Dannie is reasoning at a multi-structural level when making generalizations since, although she is recognizing sample size to be important when making an inference, she is not using the results from all the samples to make generalizations. At the time Dannie took the retention test, Task12RetT, she reasoned in a similar way she did during Task8PostT. A sample of size ten is small for her to make inference about the button, but she recognized that a sample of size 100 could be use to make inference about the button. After 1000 she was convinced the button was unfair "because of what happened the other times." Here Dannie demonstrated to pay attention to the pattern after ten, 100, and 1000 tosses using more than a single sample to make generalizations. Dannie is reasoning at a relational level.

Variability. Regarding the variability construct in the framework, Dannie paid attention to variability, but only among heads and tails in a single sample during Task1PreT. She makes no comparisons between the outcomes of the three samples (10 tosses, 100 tosses, and 1000 tosses). The same was observed during Task8PostT. Regarding this construct Dannie is reasoning at a uni-structural level recognizing variability within a sample, but not across samples in Task1PreT and Task8PostT. Dannie recognized variability within and across samples during Task12RetT. She observed the difference between the number of times the button landed ridged side up and side down within the sample of size 100, but at

the same time she noticed that the same happened within the sample of size 1000, observing variability across samples. This reasoning situates Dannie at a multi-structural level within the variability construct.

Table 4.5. Dannie's SOLO level in each task – 1st sequence

	Task1PreT	**Task8PostT**	**Task12RetT**
Generalization	Uni-structural	Multi-structural	Relational
Variability	Uni-structural	Uni-structural	Multi-structural

Dannie's 2nd sequence

The second sequence is included in Appendix C, Table C.2. This sequence asked to describe what she will do to determine the fairness of a coin (Task2PreT), a button (Task9PostT), and a bottle cap (Task13RetT), and what would give her convincing evidence.

Generalization. Dannie started the second sequence at a pre-structural level of reasoning in all three of the constructs in the framework: generalization, variability, and investigation process. In Task2PreT Dannie suggested to "try the coin out" in order to make generalizations about the fairness of the coin, reasoning at a pre-structural level in the generalization construct since she is not suggesting a size for the sample to be collected. As she moved through the instruction, she seems to have gained recognition of the importance of collecting data and sample size. After instruction, in Task9PostT, Dannie made generalizations about the fairness of the bottle cap from a sample of size 5000. She stated it

will gave her convincing evidence "if it were pretty close" but it is not clear whether she is referring to the distribution of the outcomes. The use of a large sample size, and the lack of evidence on how she will use the results to make generalizations, situates her reasoning at a multi-structural level in the generalization construct. During Task13RetT, she uses a sample of size 100 to get convincing evidence about the fairness of the button, reasoning at a multi-structural level.

Variability. Dannie "will try the coin out" and will determine the fairness of it based in "if it worked." Given the vague argument she is giving about expectations of variability, her reasoning is located at the pre-structural level in the variability construct. Dannie showed awareness of variability among the individual results within a sample size in Task9PostT since she is expecting the results to be "pretty close", reasoning at a uni-structural level. Although there is no evidence Dannie is recognizing variability across samples, Dannie's level of reasoning moved up to the multi-structural level in Task13RetT since she is expecting and recognizing variability within a sample. She also stated she will be convinced the button was more likely to land on one way "if more than 75 times it was one side" describing the variability she expected.

Investigation. In the first task, Dannie did not propose a data collection process, stating "I would try the coin out" which makes no reference to the size of the sample she will use or what she will attend to in the data generated from her trying it out. This locates her in the pre-structural level of reasoning in the investigation construct. During Task9PostT she

proposed to "flip [the bottle cap] 5000 times" and decide if it was fair based on "if it were pretty close." Since she proposed an appropriate data collection process and described how to assess the results she is on a relational level of reasoning. In Task13RetT, although she proposed to collect a sample of size 100, she will decide the button is unfair if it landed in one side "more than 75 times", reasoning at a multi-structural level in the investigation construct.

Table 4.6. Dannie's SOLO level in each task – 2nd sequence

	Task2PreT	**Task9PostT**	**Task13RetT**
Generalization	Pre-structural	Multi-structural	Multi-structural
Variability	Pre-structural	Uni-structural	Multi-structural
Investigation	Pre-structural	Relational	Multi-structural

Dannie's 3rd sequence

Three tasks were included in this sequence (refer to Appendix C, Table C.3). Task3PreI asked students to determine the number of times they needed to draw a single marble, out of a bag containing 10 marbles, in order to be very confident about how many marbles of each color are in the bag. Task 10PostI was similar to Task3PreI but using a jar of Jolly Ranchers instead of a bag of marbles. Task 5Inst asked to determine the content of a bag of marbles just knowing that it contains 12 marbles. No information about the colors was given.

Generalization. At first Dannie makes generalizations using a sample of size six, because "it is one marble over a half and that would probably tell you how much of each color would be" in the bag. In the second part of the task she proposes to make generalizations about the content of the bag using a sample of size ten "because that would tell you how many of each color are in the bag." She used a subjective strategy leading her to select small samples sizes and she did not use the distributions of the outcomes to make generalizations, locating her in the uni-structural level of reasoning in the generalization construct. Moreover, she mentioned that getting 1 blue marble and 9 red marbles after 10 tosses tells her there are 9 red marbles and 1 blue marble in the bag indicating she does not recognize the importance of sample size to make generalizations. In Task5Inst she is using several samples of a small size to make generalizations about the colors present in the bag of marbles. In this task she used the most frequent number (the mode) to decide on how many marbles of each color were in the bag (e.g., "I did this because there was most 1's in the green column"). Her prediction was there were one green marble, six yellow marbles, and five blue marbles in the bag. The true distribution was one green, five yellows, and six blues. Her reasoning is at a multi-structural level since she is using a comparison of the data across several samples to make generalizations. At the end of the sequence, Task10PostI, she generalized about the content of the jar of Jolly Ranchers using a sample of size 100 "because that equals to a whole." She also used the distribution of the outcomes to make generalizations about the content of the jar, reasoning at a multi-structural level.

Variability. Dannie does not expect to have variability among the individual results

within a single sample since she expects the number of draws to match with number of colors in the marbles in the bag, reasoning at a pre-structural level in the variability construct in Task3PreI. In Task5Inst she recognized the variability across different samples since she used the mode to make her predictions about the content of the bag. Since she is not recognizing variability within a single sample she is reasoning at a uni-structural level. In Task10PostI there is no evidence to determine Dannie recognized variability within and/or across samples, locating her at a pre-structural level.

Investigation. In the investigation construct Dannie started reasoning at a uni-structural level since she proposed to use a small sample. When asked how many times she will need to draw a single marble to be very confident about how many marbles of each color were in the bag she said "probably six because that would be one over a half of how many marbles are in the bag", in other words she propose six because it is equal to one more than half of ten, the total number of marbles in the bag (10/2 + 1 = 6). Then when the number of colors was increased from two to four she proposed a sample of size ten, using the "total weight approach". During the instructional task, Task5Inst, she uses several samples of size 12, again using the "total weight approach" since the total number of marbles in the bag was f12. At this point, she is reasoning at a multi-structural level since she is collecting more than a single sample. In Task10PostI Dannie proposed to collect a sample of size 100 because "100 is the full amount for percents… 100 would give me a lot because that equals one whole." Even though she seems to be paying attention to percentages, she did not demonstrate that reasoning when asked what it would tell her if after 100 tosses she got 83

cherries and 17 grapes. Her response was that there were 3 grapes and 7 cherries "because I use the second number, and I switched them around because there were more cherries." Her reasoning is at a multi-structural level since she is proposing a sample of size 100, but she is assessing the results inappropriately.

Table 4.7 Dannie's SOLO level in each task – 3rd sequence

	Task3PreI	**Task5Inst**	**Task10PostI**
Generalization	Uni-structural	Multi-structural	Multi-structural
Variability	Pre-structural	Uni-structural	Pre-structural
Investigation	Uni-structural	Multi-structural	Multi-structural

Dannie's 4th sequence

This sequence includes four tasks (see Appendix C, Table C.4). Task4PreI had students reasoning about the likelihood of a plastic cup to land upside down. Task6Inst presented a situation about a lake that has two types of fish in it, and they needed to estimate the probability to catch a Blue Bass. In Task7Inst students needed to determine whether or not the dice sold by a given company was fair or not, and Task11PostI was similar to Task4PreI but using a toothpaste cap instead of a plastic cup.

Generalization. Dannie started reasoning at a pre-structural level since she "wasn't sure" how many times she would need to drop the plastic cup in order to be confident about the likelihood that the cup would land upside down. Then, similar to Lara, in Task6Inst she

makes generalizations about the probability of catching a Blue Bass based on several samples of size 20 and then "took the averages and turned it into a fraction." Her estimation that "3 out of 10 catches" would be a Blue Bass was pretty close to the theoretical probability of 1 out of 3 catches. Her reasoning is at a multi-structural level.

During the Schoolopoly task, Task7Inst, Dannie (same as Lara) recommended the school "should buy dice from this company because from this evidence they are fair." She used several samples of different sizes, ranged from n=6 to n=100, and used the distribution of the outcomes to make generalizations. Her level of reasoning is at a multi-structural level, since she is not making the connection between sample size and generalizations. The dice sold by Dice R' Us were after all unfair, but with a pretty even distribution (see Table 3.3). In Task11PostT, Dannie used a sample of size 100 to make generalizations and used the distribution of the outcomes to make sense of the 100 trials. After 100 trials, she was asked what it would tell her if it only landed upside down 7 times to what she responded "it is not that likely… since I did 100 that would probably be 7 percent." She is reasoning at a multi-structural level in the generalization construct.

Variability. Reasoning about variability was not observed in Dannie's response in Task4PreI. Since the task consisted of a worksheet and a poster Dannie and Lara completed together Dannie's analysis in this construct in this particular sequence is the same as Lara's. They reasoned at a multi-structural level in Task6Inst and Task7Inst since they recognized the variability across different samples and within a single sample. In Task11PostI the

interviewer asked her what would it tell her if after 100 trials the toothpaste cap landed right side up only seven times. She answered that "since [she] did it 100 that would probably be seven percent" recognizing 7 times does not mean the probability is seven percent. She is allowing variability since she is aware that the number of times the toothpaste landed right side up are different from sample to sample. She is reasoning at a uni-structural level.

Table 4.8. Dannie's SOLO level in each task – 4th sequence

	Task4PreI	**Task6Inst**	**Task7Inst**	**Task11PostI**
Generalization	Pre-structural	Multi-structural	Multi-structural	Multi-structural
Variability	Not observed	Multi-structural	Multi-structural	Uni-structural
Investigation	Uni-structural	Multi-structural	Multi-structural	Multi-structural

Investigation. In the investigation construct, although Dannie proposed a data collection she did not offer a number of times the plastic cup should be drop and did not mentioned how she will assess the outcomes in Task4PreI. Her level of reasoning here is uni-structural. In the same way her partner Lara reasoned in Task6Inst, Dannie proposed to collect several samples and used the distribution of the outcomes to assess the results. Since the samples proposed were of small size, they are reasoning at a multi-structural level. During the Schoolopoly task, Task7Inst, Dannie proposed to collect several sample of different sizes and used the distribution of the outcomes to determine the fairness of the dice. This reasoning situates her in a multi-structural level. In Task11PostI, Dannie proposed to collect a sample of size 100 and proposed to use the empirical probability distribution to

assess the results. The interviewer asked her what it would tell her if after 100 tosses the toothpaste cup landed on one side 7 times. She answered that "since [she] did it 100 that would probably be seven percent." She is reasoning at a multi-structural level in the investigation construct.

Dannie demonstrated an increase in her levels of reasoning in all four sequences in the generalization construct. Her level of reasoning seems to increase in the variability construct during the first two sequences, whereas her reasoning went back to the lowest level in the 3rd sequence. She also demonstrated to increase her level of reasoning in the investigation construct reasoning at a multi-structural level by the end of the 3rd and 4th sequence. During the 2nd sequence, although she reasoned at a relational level in Task9PostT, her reasoning was found to be at a multi-structural level by the end of the sequence.

In summary, Lara and Dannie, the high-average scoring group, showed to increase their level of reasoning in all three construct in almost all four sequences. Table 4.9 shows Lara's and Dannie's levels of reasoning. During the 1st sequence they both showed a growth on their reasoning levels. Although Lara and Dannie demonstrated to have a growth in their levels of reasoning in the generalization and variability constructs in the second sequence, they both showed decay in their levels of reasoning in the investigation construct (indicated by bolded letters). Dannie showed to go back to a pre-structural level of reasoning in the variability construct in the 3rd sequence, whereas Lara showed a constant increase in all three constructs.

During the fourth sequence both, Lara and Dannie, demonstrated to reason at a higher level from the pre task (Task4PreI) to the instructional task (Task6Inst) and then stayed constant in their level of reasoning through the end of the sequence in the generalization and investigation construct. Their levels of reasoning decay from the instructional tasks (Task7Inst) to the post task (Task11PostI) in the variability construct, but if only individual tasks (Task4PreI and Task11PostI) are considered, they both increased their level of reasoning from the pre task (Task4PreI) to the post task (Task11PostI).

Table 4.9 Lara's and Dannie's SOLO levels in all four sequences

Construct	Sequence	1st			2nd			3rd			4th			
	Task	Task1PreT	Task8PostT	Task12RefT	Task2PreT	Task9PostT	Task13RefT	Task3PreI	Task5Inst	Task10PostI	Task4PreI	Task6Inst	Task7Inst	Task11PostI
G	Lara	P	M	M	P	M	M	P	U	M	U	M	M	M
	Dannie	U	M	R	P	M	M	U	M	M	P	M	M	M
V	Lara	P	U	U	P	U	U	P	U	U	P	M	M	U
	Dannie	U	U	M	P	U	M	**P**	**M**	**P**	N/O	M	M	U
I	Lara				**M**	**M**	**U**	P	M	R	U	M	M	M
	Dannie				**P**	**R**	**M**	U	M	M	U	M	M	M

CHAPTER 5

RESULTS: *Manuel & Brandon*

There were four sequences used to analyze students' changes in probabilistic inference. Within this chapter there are two sections, each describing the reasoning of each student individually. Each section is divided into four sub-sections which describes a student's performance within each sequence. This chapter describes Manuel's probabilistic inference changes in the first section and Brandon's probabilistic inferences changes in the second section. Recall that Manuel and Brandon are the average scoring group. Tables of the frameworks are included in Appendix B, tables B.1 through B.5 and the tables of the sequences could be found in Appendix C, tables C.1 through C.4.

The case of Manuel

Manuel's 1st sequence

The first sequence is included in Appendix C, Table C.1. This sequence includes three tasks where students have to reason about the fairness of a coin, a bottle cap, and a shaped button, respectively. Only two of the three construct were used to analyze changes in probabilistic reasoning in this sequence: generalization and variability. The tasks are included in Appendix A.

Generalization. Manuel generalized the coin was unfair stating that "it should had been 50-50" after 100 tosses, and "it should had been 500-500" after 1000 tosses. His

reasoning is at a uni-structural level since he did not recognize sample size as an important factor to make generalizations. Although he recognized the difference between the number of heads and the number of tails within each individual sample he generalized about the fairness of the coin even with the sample of size 10. His level of reasoning in Task8PostT and Task12RetT was identified as a pre-structural since he is using the physical appearance of the bottle cap and the shaped button to make generalization. In Task8PostT he generalized the bottle cap was unfair "because one side may be heavier", paying no attention to the distribution of the outcomes either after 10, 100, and 1000 tosses. He also generalized the button in Task12RetT was unfair "because it is ridged on one side" and that made the button "very mobile on ridge side."

Table 5.1. Manuel's SOLO level in each task – 1st sequence

	Task1PreT	**Task8PostT**	**Task12RetT**
Generalization	Uni-structural	Pre-structural	Pre-structural
Variability	Uni-structural	Pre-structural	Pre-structural

Variability. At first, in Task1PreT, he noticed the difference between the number of heads and the number of tails within each sample, but he did not notice the pattern among the three samples, reasoning at a uni-structural level in the variability construct. After that, in Task8PostT and Task12RetT, he did not pay attention to the variability among individual results within a sample or the variability across samples, reasoning at a pre-structural level.

Manuel's 2nd sequence

This sequence asked to describe what he will do to determine the fairness of a coin (Task2PreT), a button (Task9PostT), and a bottle cap (Task13RetT), and what would give him convincing evidence.

Generalization. In all three of the tasks of the second sequence Manuel generalized about the fairness of the object (the coin, the bottle cap, and the shaped button) using a single sample, with the difference that he increased the size of that sample from task to task. His reasoning is located at a uni-structural level in Task2PreT since he is using a small sample (n=10) to make generalizations about the fairness of the coin. In Task9PostT he is still reasoning at a uni-structural level, generalizing using a sample of size 30 and in Task12RetT his level of reasoning increased to multi-structural since he specified he will "flipped [the button] 50 times and saw the outcomes" to decide about the fairness of the button.

Variability. Manuel's level of reasoning in the variability construct started and ended at a pre-structural level. From his answers to Task2PreT, Task9PostT, and Task13RetT, Manuel does not show that he recognizes variability within the sample to be collected.

Investigation. Manuel's investigation reasoning in all three tasks in this sequence stayed at a uni-structural level. In Task2PreT he started proposing to collect a sample of size 10, and to observe "the number of times [he] got something." Since Manuel is proposing a small sample size and was not specific about how he will assess the results, his reasoning is at a uni-structural level. In Task9PostT he increased the sample size, from n=10 to n=30,

but did not described how he will assess the results. He stated he will "feel comfortable with [his] answer" if he flip the bottle cap 30 times. His level of reasoning at this point is uni-structural since he is not describing how he will assess the results. By the end he proposed to collect a sample of size 50 and to "saw the outcomes" to decide whether or not the shaped button was fair. Here he did not specify how he will use the distribution of the outcomes to assess the results, reasoning is at a uni-structural level since he just mentioned he will see the outcomes to be convinced.

Table 5.2. Manuel's SOLO level in each task – 2nd sequence

	Task2PreT	**Task9PostT**	**Task13RetT**
Generalization	Uni-structural	Uni-structural	Multi-structural
Variability	Pre-structural	Pre-structural	Pre-structural
Investigation	Uni-structural	Uni-structural	Uni-structural

Manuel's 3rd sequence

Three tasks were included in this sequence. Task3PreI asked students to determine the number of times they needed to draw a single marble, out of a bag containing 10 marbles, in order to be very confident about how many marbles of each color are in the bag. Task 10PostI was similar to Task3PreI but using a jar of Jolly Ranchers instead of a bag of marbles. Task 5Inst asked to determine the content of a bag of marbles just knowing that it contains 12 marbles. No information about the colors was given.

Generalization. Manuel started reasoning at a pre-structural level in the

generalization construct since he did not predict about the content of the bag of marbles. In Task5Inst he uses several samples of a small size (n=12) and somehow uses the distribution of the outcomes to make generalizations about the content of the bag his partner Brandon created, locating his level of reasoning during this task at a multi-structural level in the generalization construct. He predicted there were 4 red, 6 black, and 2 blue marbles "cause they seemed more practical." Those numbers he selected are not the more often number in the column of red, black and blue marbles neither the average of each column, which makes unclear how he used the data collected and what he meant by "more practical." During the first part of Task10PostT Manuel proposes to make a generalization about the content of the jar of Jolly Ranchers using a sample of size 30. In the second part of the task, when the number of flavors in the jar was increased from two to four, he increased the sample size from 30 to 100 stating that 100 " was a big number." Here he recognized it seems to be necessary to increase the number of the sample size when the number of flavors was increased. He makes generalizations about the content of the jar using the distribution of the outcomes either after 30 and after 100 trials. Even that his initial sample was small, his level of reasoning is at a multi-structural level since he used the distribution of the outcomes to make generalizations. Manuel was ask what does the data tells him if after 100 tosses he got 17 grapes, 13 cherries, 27 lemons, and 43 green apples. He answered:

> "Because grapes and cherries are so close together I'll say they have the same amount… so I'll say there are 2 of each… and then green apples because it has a very big number it will have about 4. This two will have 1 each, lemons will have 3 and green apple will have 6. I

mean 5, 3, and 1 grape and 1 cherry"

Variability. In Task3PreI Manuel does not expected variability among individual results within a single sample since he assumes the sample of size 10 was representative of the underlying populations of marbles inside the bag, reasoning at a pre-structural level in the variability construct. Since it is not clear to the author what Manuel meant by "more practical" it is hard to determine whether he recognized variability within each sample. Manuel does recognize variability across different samples since he seemed to be paying attention to the outcomes within each color. His level of reasoning during this task is uni-structural. After instruction, Task10PostI, Manuel recognized variability within a sample, as described at the end of the previous paragraph. He is reasoning at a uni-structural level during this task.

Investigation. In the investigation construct Manuel started reasoning at a pre-structural level in Task3PreI since he does not propose any data collection process. He stated that "it does not matter how many times you do it as long as you find every single marble." He thinks that in order to determine how many marbles of each color are in the bag he needs to "know the percentages of it… it would depend on the percentage of the marbles." What he possibly meant here is that he would need to know the percentages of the different colors in order to know the content of the bag, not to determine how many times he will need to draw a single marble. For example, if there is a bag containing 12 marbles and he knows there is a 25% chance to get a red marble, 67% chance to get a blue marble, and 1% chance to get a

white marble, then he knows there are 3 red, 8 blue, and 1 white marble in the bag. In Task5Inst his level of reasoning is at a multi-structural since he proposed to collect several samples of a size 12, although following the "total weight approach." He did not describe how he determined there were 4 reds, 6 blacks, and 2 blues in the bag of marbles and his only comment was "cause they all seemed more practical." During Task10PostI he proposed to collect first, a sample of size 30, and then a sample of size 100. He also used the distribution of the outcomes to make sense of the data collected and assess the results. His reasoning locates him at a relational level in the investigation construct since he proposed an appropriate data collection and reasoned appropriately on how to assess the results.

Table 5.3. Manuel's SOLO level in each task – 3rd sequence

	Task3PreI	**Task5Inst**	**Task10PostI**
Generalization	Pre-structural	Multi-structural	Multi-structural
Variability	Pre-structural	Uni-structural	Uni-structural
Investigation	Pre-structural	Multi-structural	Relational

Manuel's 4th sequence

There are four tasks in this sequence. Task4PreI had students reasoning about the likelihood of a plastic cup to land upside down. Task6Inst presented a situation about a lake that has two types of fish in it, and they needed to estimate the probability to catch a Blue Bass. In Task7Inst students needed to determine whether or not the dice sold by a given company was fair or not, and Task11PostI was similar to Task4PreI but using a toothpaste

cap instead of a plastic cup.

Generalization. In Task4PreI, Manuel started saying "[he was] not really sure" about how many times he will need to drop the plastic cup stating "it just depend on the percentages." He said that "you could drop it all day and not be confident about the guess." His reasoning here is at a uni-structural level since he is not using a sample of a determined size to estimate the likelihood that the plastic cup would land upside down. During Task6Inst, Manuel had to determine the probability that a fish caught in a lake was a Blue Bass. He makes generalizations about the probability based on several samples of different sizes (n=500, 3, and 30) and used the distribution of the outcomes to estimate the probability of catching a Blue Bass. How he use the distribution of the outcomes to estimate the probability is not clear on the worksheet, but he estimated there were 1 Blue Bass per 2 Green Gill, which was an exact estimation of the theoretical probability of 1 out of 3 catches. His reasoning here is at multi-structural level because although he uses several samples, most of them had a pretty small size (n=3 for 5 of the 8 samples), and it is not clear how he used the distribution of the outcomes to estimate the probability of catching a Blue Bass or a Green Gill.

During the Schoolopoly task, Task7Inst, Manuel investigated the fairness of the dice sold by High Rollers, Inc. with his partner Brandon. Manuel (same as Brandon) did not recommend the school to buy dice produced by this company because "the info [they] collected proved that High Rollers was unfair." At first they collected a sample of size 50 and

used the distribution of the outcomes to generalize the dice were fair. Then they increased the sample size and they ran 4 samples of size 500 for a total of 200 trials. Using the distribution of the outcomes they generalized they had changed their mind and that "the dice are soooooo sooooo soooo unfair." They continued running samples of size 500 until they reached 6000 trials, confirming what they found before that "it is still unfair because of the evidence of the percents on the table." His level of reasoning is at a relational level since they make generalizations based on large sample and used the distribution of the outcomes to make sense of their results. They noticed the importance of large samples, ignoring the results after 50 trials and moving on collecting more data. He used the empirical probability distribution after 6000 trials as his estimations of the theoretical probability of each outcome. The dice sold by High Rollins, Inc. were unfair, and Manuel predicted its distribution almost perfectly (see Table 3.3). In Task11PostI Manuel estimated the likelihood that the toothpaste cap would land upside down using a small sample (n=20) and he stated that if after 20 times it never landed upside down then he will "just say it is unfair." His level of reasoning during this task it as a uni-structural since he is not using the distribution of the outcomes to estimate the likelihood that the toothpaste cap landed upside down.

Variability. Reasoning about variability was not observed in Manuel's answer in Task4PreI. In Task6Inst Manuel (same as Brandon) recognized variability across samples as he noticed the difference in the outcomes he got in the different samples. He used the mode to decide on the proportions of Blue Bass and Green Gill. He also noticed the difference among the number of times he got a Blue Bass and the number of times he got a Green Gill

within every single sample since he recognized there were more Green Gill fishes in the lake. He is reasoning at a multi-structural level during this task. In Task7Inst, he recognized variability within and across samples. He expected too much variability stating the dice were fair even after getting 9 ones, 6 twos, 4 threes, 11 fours, 9 fives, and 11 sixes out of 50 trials. Then, after 2000 trials he recognized the distribution of the outcomes were too variable to decide the dice were fair. His level of reasoning here is multi-structural. In Task11PostI Manuel did not recognize variability, expecting to get all three possible outcomes from a sample of a small size (n=20). His is reasoning at a pre-structural level in this task.

Table 5.4. Manuel's SOLO level in each task – 4th sequence

	Task4PreI	**Task6Inst**	**Task7Inst**	**Task11PostI**
Generalization	Uni-structural	Multi-structural	Relational	Uni-structural
Variability	Not observed	Multi-structural	Multi-structural	Pre-structural
Investigation	Pre-structural	Multi-structural	Relational	Multi-structural

Investigation. Manuel proposed an inappropriate data collection in Task4PreI suggesting that collecting data will not be useful to estimate the likelihood the plastic cup would land upside down, and does not mention how he will assess the results, locating his reasoning at a pre-structural level. In Task6Inst he collected several samples of small sizes and to use the distribution of the outcomes to assess the results, reasoning at a multi-structural level. In Task7Inst Manuel proposed to collect several large samples and to use the distribution of the outcomes to assess the results. He recognized the importance of sample

size in order to make generalization, situating him at a relational level. In the last task of task of the sequence Manuel proposed to collect a small sample (n=20) and recognized the outcomes will tell him some information about the likelihood of the toothpaste cap to land upside down. He is proposing to use the distributions of the outcomes to assess the results and is proposing to collect data in order to be convincing about the likelihood of the toothpaste cap to land upside down, but is proposing a small sample, reasoning at a multi-structural level.

In general, Manuel's levels of reasoning were very inconsistent throughout the sequences. He started reasoning at a uni-structural level in the 1st sequence and demonstrated to be reasoning at a lower level by the end of the sequence in the generalization and variability constructs. His level of reasoning stayed constant during the 2nd sequence, with the exception of the generalization construct. Manuel seemed to be reasoning at the highest level of reasoning, the relational level, only once in the tasks he completed individually, If the instructional tasks are removed from the 3rd and 4th sequence it could be said that Manuel's level of reasoning increased in the generalization, variability, and investigation constructs.

The case of Brandon

Brandon's 1st sequence

This sequence includes three tasks where students have to reason about the fairness of a coin, a bottle cap, and a shaped button, respectively. Only two of the three construct were

used to analyze changes in probabilistic reasoning in this sequence: generalization and variability.

Generalization. Brandon generalized about the fairness of the coin in Task1PreT based in his previous experiences with coins, reasoning at a pre-structural level. After 10, 100, and 1000 tosses he concluded the coin was fair "because it has heads and tails." In Task8PostT he noticed the difference between the number of times the bottle cap landed up and down, but generalized the bottle cap was fair saying it was "just luck." Brandon stayed at a pre-structural level of reasoning and moved to a relational level of reasoning in Task12RetT by noticing that a sample of size 10 had "not enough info" to make generalizations, "still not sure" after 100 tosses, and generalize the button was unfair after 1000 tosses because it "weights more on flat side."

Variability. On Task1PreT Brandon did not recognize variability and in Task8PostT he attributed variability to luck, indicating a pre-structural level of reasoning. In Task12RetT, although after 1000 tosses he still attributed variability to luck, he recognized the variability among the individual results within the 1000 tosses sample, increasing his level of reasoning to uni-structural.

Table 5.5. Brandon's SOLO level in each task – 1st sequence

	Task1PreT	**Task8PostT**	**Task12RetT**
Generalization	Pre-structural	Pre-structural	Relational
Variability	Pre-structural	Pre-structural	Uni-structural

Brandon's 2nd sequence

This sequence asked to describe what he will do to determine the fairness of a coin (Task2PreT), a button (Task9PostT), and a bottle cap (Task13RetT), and what would give him convincing evidence.

Generalization. Brandon reasoned at a pre-structural level in the first two tasks: Task1PreT and Task8PostT. His generalization was based on previous experience with coins, believing that all coins are fair if "it has heads and tails" and bottle caps are fair if "show both sides." In Task13RetT he increased the size of his sample to 100, since he did not mention how to make sense of those 100 trials he is reasoning at a uni-structural level in the generalization construct since he is using a relatively large sample.

Variability. Reasoning about variability was not observed in any of the tasks. Brandon did not appear to consider variability either within a single sample or across samples, reasoning at a pre-structural level in all three tasks of the sequence.

Table 5.6. Brandon's SOLO level in each task – 2nd sequence

	Task2PreT	**Task9PostT**	**Task13RetT**
Generalization	Pre-structural	Pre-structural	Uni-structural
Variability	Pre-structural	Pre-structural	Pre-structural
Investigation	Pre-structural	Pre-structural	Uni-structural

Investigation. Brandon did not propose a data collection process to determine the

fairness of the coin in Task2PreT and the fairness of the bottle cap in Task9PostT reasoning at a pre-structural level. By the time he took the retention test and completed Task13RetT, Brandon proposed to collect a sample of size 100, but he did not describe how to assess the results, reasoning at a uni-structural level since he is focusing on a single aspect of the situation presented.

Brandon's 3rd sequence

Three tasks were included in this sequence. Task3PreI asked students to determine the number of times they needed to draw a single marble, out of a bag containing 10 marbles, in order to be very confident about how many marbles of each color are in the bag. Task 10PostI was similar to Task3PreI but using a jar of Jolly Ranchers instead of a bag of marbles. Task 5Inst asked to determine the content of a bag of marbles just knowing that it contains 12 marbles. No information about the colors was given.

Generalization. Bandon stated he will be very confident about how many marbles of each color are in the bag using a sample of size 30, either if there were two or four colors in the bag. Since he used a single small sample to make generalizations about the content of the bag, his reasoning is at the uni-structural level in the generalization construct. There is no evidence about whether or not he will use the distribution of the outcomes to make generalizations, but he stated that if there were 1 blue and 9 red after 100 trials, the only information he would have about the content of the bag was that there were "probably more reds."

In Task5Inst Brandon's level of reasoning was at a multi-structural level since he used eight samples of size 12 to make a generalization about the bag of marbles that his partner Manuel created. His prediction were pretty close to the actual content of the bag and he stated that he based his prediction "on the trials" without giving any particular description about how he used the trials to make his final predictions. In Task10PostT, Brandon uses a sample of size 100 to make generalizations about the content of the jar of Jolly Ranchers when there were two colors in the jar, and a sample of size 200 when there were four colors. There is no evidence he will use the distribution of the outcomes to make generalizations about the content of the bag his partner Manuel created. His reasoning at this point is at a uni-structural level since his generalizations are based on a single sample of a large sample size.

Variability. In the variability construct Brandon seemed to recognize variability within a single sample, locating his reasoning at a uni-structural level in Task3PreI. When asked what would tell him if after 10 trials he gets one blue and 9 red marbles he answered "actually not much really, but probably there are maybe more red." In this line, Brandon is recognizing there is variability within a sample. In Task5Inst he is reasoning at a uni-structural level since he recognized variability across different samples, but there is no evidence he recognized variability within a single sample. In Task10PostT there is no evidence he recognized either variability within or across samples, situating his reasoning at a pre-structural level.

Table 5.7. Brandon's SOLO level in each task – 3rd sequence

	Task3PreI	**Task5Inst**	**Task10PostI**
Generalization	Uni-structural	Multi-structural	Uni-structural
Variability	Uni-structural	Uni-structural	Pre-structural
Investigation	Uni-structural	Multi-structural	Uni-structural

Investigation. Brandon proposed to collect a sample of size 30 in Task3PreI, but did not propose how he will assess the results. His reasoning at the beginning of this sequence is at a uni-structural level. During instruction, Task5Inst, he collected several samples and proposed how to assess the results "base on the trials." During this task his level of reasoning is multi-structural since the samples he proposed are of small size, but reasoned appropriately on how to assess the results using the distribution of the outcomes. In Task10PostT Brandon reasoning level moved back to uni-structural since, although he proposed an appropriate data collection, he did not showed evidence about how he will assess the results.

Brandon's 4th sequence

This sequence has two tasks from interviews and two tasks from the instruction. Task4PreI had students reasoning about the likelihood of a plastic cup to land upside down. Task6Inst presented a situation about a lake that has two types of fish in it, and they needed to estimate the probability to catch a Blue Bass. In Task7Inst students needed to determine whether or not the dice sold by a given company was fair or not, and Task11PostI was

similar to Task4PreI but using a toothpaste cap instead of a plastic cup.

Generalization. In Task4PreI, Brandon started reasoning at a pre-structural level in the generalization construct. Even though he said he will drop the plastic cup 12 times, he mentioned that "all [outcomes] has an equal chance [and] it really depends on how your drop it." During Task6Inst, Brandon had to determine the probability that a fish caught in a lake was a Blue Bass. He makes generalizations about the probability based on several samples of different sizes and used the distribution of the outcomes to estimate the probability of catching a Blue Bass. He used mode to estimate there were 1 Blue Bass per 2 Green Gill, which was an exact estimation of the theoretical probability of 1 out of 3 catches. His reasoning here is at multi-structural level.

During the Schoolopoly task, Task7Inst, Brandon investigated the fairness of the dice sold by High Rollers, Inc. with his partner Manuel. Brandon (same as Manuel) did not recommend the school to buy dice produced by this company because "the info [they] collected proved that High Rollers was unfair." The details about their analysis were presented above in Manuel's analysis. His level of reasoning is at a relational level since they made generalizations based on a large sample and used the distribution of the outcomes to make sense of their results. They noticed the importance of large samples, ignoring the results after 50 trials and moving on collecting more data. Task11PostI Brandon estimated the likelihood that the toothpaste cap would land upside down using a reasonable sample size (n=100) and he stated that "the side will have a greater chance because it has a bigger area of

falling." Although he used a sample to estimate the likelihood, his generalizations were not based on the distribution of the outcomes but in the physical appearance of the toothpaste cup, observing that the side had a greater area so it would have a greater chance. His level of reasoning during this task it as a uni-structural since he does not seems to recognize he could use the distribution of the outcomes to estimate the likelihood that the toothpaste cap landed upside down.

Variability. Brandon's level of reasoning about variability was not observed in Task4PreI. Since Task6Inst and Task7Inst consisted of a worksheet and a poster, respectively, Manuel and Brandon completed together, Brandon's analysis in this construct in this particular sequence is the same as Manuel's. They started reasoning at uni-structural level, in Task6Inst, as they recognized the variability across different samples and moved to a multi-structural level, in Task7Inst, as they recognized variability within and across samples. Brandon's level of reasoning about variability was not observed in Task11PostI.

Investigation. Bandon proposed an inappropriate data collection (n=12) in Task4PreI and does not mentioned how he will assess the results, locating his reasoning at a pre-structural level. In Task6Inst he collected several samples of small sizes and to use the distribution of the outcomes to assess the results, reasoning at a multi-structural level. In Task7Inst Brandon collected several large samples and to use the distribution of the outcomes to assess the results. He recognized the importance of sample size in order to make generalization, situating him at a relational level. In the last task of task of the sequence

Brandon proposed to collect a sample of size 100 but did not mentioned how he will analyze the results obtained from that sample, reasoning at a uni-structural level.

Table 5.8. Brandon's SOLO level in each task – 4th sequence

	Task4PreI	**Task6Inst**	**Task7Inst**	**Task11PostI**
Generalization	Pre-structural	Multi-structural	Relational	Uni-structural
Variability	Not observed	Uni-structural	Multi-structural	Not observed
Investigation	Pre-structural	Multi-structural	Relational	Uni-structural

Brandon seemed to increase his levels of reasoning during the 1st and 2nd sequence. During the 3rd sequence, although his level of reasoning increased to a multi-structural level in the generalization construct his level of reasoning in the variability and investigation constructs showed to decrease by the end of the sequence. If the instructional tasks, Task6Inst and Task7Inst, are not taking into account in the 4th sequence, Brandon's level of reasoning in the generalization and investigation constructs increase.

In summary, Manuel and Brandon, the average scoring group, showed to increase their level of reasoning in the generalization and investigation constructs, whereas in the variability construct they seem to be reasoning, mostly, at a pre-structural and uni-structural level, the two lower levels of reasoning. Table 4.10 shows Manuel's and Brandon's levels of reasoning. During the first sequence Brandon showed an increase in his level of reasoning whereas Manuel demonstrated to be inconsistent in his level of reasoning, reasoning a higher

level at the beginning of the sequence and at a lower level by the end of the first sequence. During the second sequence only Manuel demonstrated to be reasoning at a multi-structural level, but in most of the instances showed in Table 4.10 they both were reasoning at a pre-structural and uni-structural level. During the third sequence, Brandon demonstrated inconsistencies in his reasoning in the variability and investigation constructs (refer to cells with letters on bold). Manuel's level of reasoning achieved the relational level in the investigation construct. Their levels of reasoning increased from Task4PreI to Task6Inst in the fourth sequence, but showed to decrease by the end of the sequence. If only individual tasks (Task4PreI and Task11PostI) are considered, they both increased their level of reasoning from the pre task (Task4PreI) to the post task (Task11PostI).

Table 5. 9. Manuel's and Brandon's SOLO levels in all four sequences

Construct	Sequence	1st			2nd			3rd			4th			
	Task	Task1PreT	Task8PostT	Task12RetT	Task2PreT	Task9PostT	Task13RetT	Task3PreI	Task5Inst	Task10PostI	Task4PreI	Task6Inst	Task7Inst	Task11PostI
G	Manuel	**U**	**P**	**P**	U	U	M	P	M	M	U	M	R	U
	Brandon	P	P	R	P	P	U	U	M	M	P	M	R	U
V	Manuel	**U**	**P**	**P**	P	P	P	P	U	U	N/O	M	M	P
	Brandon	P	P	U	P	P	P	**U**	**U**	**P**	N/O	M	M	N/O
I	Manuel				U	U	U	P	M	R	P	M	R	M
	Brandon				P	P	U	**U**	**M**	**U**	P	M	R	U

CHAPTER 6

RESULTS: *Greg & Jasyn*

There were four sequences used to analyze students' changes in probabilistic inference. Within this chapter there are two sections, each describing the reasoning of each student individually. Each section is divided into four sub-sections which describes a student's performance within each sequence. This chapter describes Greg's probabilistic inference changes in the first section and Jasyn's probabilistic inferences changes in the second section. Recall that Greg and Jasyn are the low scoring group. Tables of the frameworks are included in Appendix B, tables B.1 through B.5 and the tables of the sequences could be found in Appendix C, tables C.1 through C.4.

The case of Greg

Greg's 1st sequence

This sequence includes three tasks where students have to reason about the fairness of a coin, a bottle cap, and a shaped button, respectively. Only two of the three construct were used to analyze changes in probabilistic reasoning in this sequence: generalization and variability.

Generalization. Greg's generalization reasoning changes from uni-structural to multi-structural. In Task1PreT he generalized the coin was unfair after 10 and 100 tosses stating there were a "uneven number of flips" referring to the difference between the number of tails

and heads, reasoning at a uni-structural level since he is not paying attention to sample size. But after 1000 tosses, with 643 tails and 357 heads, he generalized that the coin was fair because "it can always change." His expression seems to indicate he expect all coins to be fair. This was confirmed with his comment in Task2PreT, when he said that "[he] has use it" indicating he beliefs all coins are fair. The same reasoning was observed in Task8PostT where he generalized, after 10, 100, and 1000 tosses that the coin was unfair because "it is not even." He is not paying attention to sample size, generalizing about the fairness of the bottle cap even when a sample of size 10. By the end of the sequence, in Task12RetT, he recognized the importance of sample size when he did not generalize after 10 tosses because he needed "more data." After 100 and 1000 tosses he generalized the button was more likely to land with the ridge side up because "ridges up [were] majority." His reasoning here is at a multi-structural level.

Variability. Since Greg was paying attention to the difference between the number of heads and tails in Task1PreT, the number of times the bottle cap landed up and down in Task8PostT, and the number of times the shaped button landed with the ridged side up and down he is reasoning at a uni-structural level from the beginning to the end of the sequence. He is recognizing the variability among the outcomes within each single sample, but is not recognizing variability across samples.

The tasks in this sequence asked Greg to describe what he would do to determine the fairness of a coin (Task2PreT), a button (Task9PostT), and a bottle cap (Task13RetT), and

what would give him convincing evidence. Tasks are included in Appendix A.

Table 6.1. Greg's SOLO level in each task – 1st sequence

	Task1PreT	**Task8PostT**	**Task12RetT**
Generalization	Uni-structural	Uni-structural	Multi-structural
Variability	Uni-structural	Uni-structural	Uni-structural

Greg's 2nd sequence

Generalization. In order to make a generalization about an old coin found on the side of the street, Greg said it is fair because "[he] has use it" indicating his belief that all coins are fair as discussed in the previous section. Here, Task2PreT, his reasoning is at a pre-structural level since he is generalizing based in his previous experiences with coins. In the following tasks, Task9PostT and Task13RetT, he used a sample of size 1000 to make generalizations, moving him to a multi-structural level of reasoning.

Variability. In Task2PreT Greg did not expect variability from tossing a coin since he believes all coins are fair, reasoning at a pre-structural level in the variability construct. Since the only aspect he mentioned in Task9PostT and Task13RetT was the number of times he would flip the bottle cap and the shaped button, respectively, there is no evidence whether he recognizes variability within and/or across samples, reasoning at a pre-structural level as well.

Investigation. In the investigation construct Greg started reasoning at a pre-structural

level. He did not propose a data collection process indicating he already knew the coin was fair. In the following tasks, Task9PostT and Task13RetT, his reasoning increased one level. He proposed a data collection process which includes a large sample size (n=1000), but did not describe how the results will be assessed. Following the framework his reasoning is at a uni-structural level.

Table 6.2. Greg's SOLO level in each task – 2nd sequence

	Task2PreT	**Task9PostT**	**Task13RetT**
Generalization	Pre-structural	Multi-structural	Multi-structural
Variability	Pre-structural	Pre-structural	Pre-structural
Investigation	Pre-structural	Uni-structural	Uni-structural

Greg's 3rd sequence

Three tasks were included in this sequence. Task3PreI asked students to determine the number of times they needed to draw a single marble, out of a bag containing 10 marbles, in order to be very confident about how many marbles of each color are in the bag. Task 10PostI was similar to Task3PreI but using a jar of Jolly Ranchers instead of a bag of marbles. Task 5Inst asked to determine the content of a bag of marbles just knowing that it contains 12 marbles. No information about the colors was given.

Generalization. Greg started reasoning at a multi-structural level in the generalization construct. He had convincing evidence about the content of the bag of marbles

after drawing a single marble, without replacement, 100 times. The interviewer did not give him results from a large sample to motivate him generalize about the content of the bag from the sample, but after being asked what does it tell him if after 10 trials he got 1 blue and 9 red he responded that "probably there were more red marbles, but it does not necessarily mean there are only one blue marble and nine red marbles" recognizing he cannot use results from small samples to make generalizations.

During the instruction, in Task5Inst, he still reasoning at a multi-structural level since he used several samples, although of size 12, to make generalizations about the content of the bag of marbles his partner Jasyn created. He also used the distribution of the outcomes to make generalizations stating "the colors show [the results]". Although he seemed to be using the mode, he did not specify how he used the distribution to make the generalizations, but his final prediction (4 yellow, 3 green, and 2 white) was a pretty good estimation of the true distribution (6 yellow, 4 green, and 2 white). At the end of the sequence, in Task10PostI, Greg make generalizations about the content of the jar of Jolly Ranchers using a sample of size 400. The interviewer asked him what does it tells him if after 400 tosses he got 40 green apples, 80 cherries, 80 grapes, and 200 lemons. He used the distribution of the outcomes and answered "there are lots of lemons in there, maybe 6 or 7… no I'll say 5 because 200 is half of 400, then 2 grape and 2 cherries, then maybe only 1 green apple." His level of reasoning is at a relational level since he used a sample of large size and used the distribution of the outcomes to make generalizations.

Variability. Greg recognized variability within samples in Task3PreI, but did not how evidence of recognition of variability across samples, locating him at a uni-structural level in the variability construct. He recognized that there are more cherries than grapes if from 100 tosses he got 75 cherries and 25 grapes. Although it is hard to determine whether he recognized variability within and across samples in Task5Inst, Greg's predictions demonstrated he could be recognizing both types of variability which locates him at a multi-structural level. His predictions of 2 white, 3 green, and 4 yellow coincide with the results of most of the samples, where whites got the less and yellows got the most, noticing the variability within samples. He recognized variability across sample since he wrote "the colors show [the results]" demonstrating she was looking at the columns which indicate he was observing across samples. In Task10PostT he seemed to recognize variability within a sample since he observed that obtaining 75 cherries and 25 grapes would mean "there are more cherries than grapes." There is no evidence he recognized variability across samples, locating his level of reasoning at a uni-structural.

Investigation. In the investigation construct, Greg started reasoning at a uni-structural level since he proposed a data collection, but did not proposed how he would assess the results. He proposed to draw a single marble 100 times. During instruction, Task5Inst, he proposed a data collection that consisted of eight samples of size 12. Then he described he used the distribution of the outcomes to assess the results, reasoning at a multi-structural level. In Task10PostT, after the instruction, he propose and appropriate data

collection (using a sample of size 400) and used the distribution of the outcomes to assess the results obtained from the sample. His level of reasoning during this task is relational since he is used an appropriate data collection and he used the distribution of the outcomes to assess the results.

Table 6.3 Greg's SOLO level in each task – 3rd sequence

	Task3PreI	**Task5Inst**	**Task10PostI**
Generalization	Multi-structural	Multi-structural	Relational
Variability	Uni-structural	Multi-structural	Uni-structural
Investigation	Uni-structural	Multi-structural	Relational

Greg's 4th sequence

Task4PreI had students reasoning about the likelihood of a plastic cup to land upside down. Task6Inst presented a situation about a lake that has two types of fish in it, and they needed to estimate the probability to catch a Blue Bass. In Task7Inst students needed to determine whether or not the dice sold by a given company was fair or not, and Task11PostI was similar to Task4PreI but using a toothpaste cap instead of a plastic cup.

Generalization. In Task4PreI Greg makes generalizations about the likelihood of a plastic cup to land upside down with a small size (n=6, 9, or 12) but does not specify how he will make sense of those 12 outcomes, reasoning at a uni-structural level. During Task6Inst, Greg had to determine the probability that a fish caught in a lake was a Blue Bass. He makes

generalizations about the probability based on six samples of size 30 each. In order to estimate the probability that a fish caught is a Blue Bass or a Green Gill he uses the distribution of the outcomes, calculating the average and converting them to a percent. His estimation of 70% probability to catch a Green Gill and 30% probability to catch a Blue Bass was pretty close to the true underlying population of 1 Blue Bass per 2 Green Gills. His level of reasoning in this task is a multi-structural since he is not considering sample size.

In Task7Inst, the Schoolopoly task, Greg investigated the fairness of the dice sold by Slice -N- Dice with his partner Jasyn. Greg (same as Jasyn) recommended the school "should probably not buy a dice from this company because there is barely any five." He makes this generalization based on results from three samples (n=42, 561, 100). After 42 trials he become suspicious the dice were unfair stating "we think this company is unfair", and then after 561 trials they become more convinced and said that "[the] company seems unfair [because] 5 is coming up a very little amount." They ran 100 trials and used the counts on the distribution of the outcomes instead of the percentages or the proportions to estimate the theoretical probability of each outcome. The dice sold by Slice -N- Dice were unfair with a really low probability of getting a 5. The theoretical probability distribution is presented in Table 3.3. His level during this task is multi-structural since he used samples of appropriate size, but although he used the distribution of the outcomes from several samples to make generalizations he did not pay attention to the percentages or the proportions, but to the counts.

In Task11PostI his generalization about the likelihood of the toothpaste cap to land upside down were based on a single sample of size 300. Although he seemed to be using the distribution of the results, he was doing it inappropriately. Greg was asked how he would estimate the probability that the toothpaste cap would land upside down if after 300 times the toothpaste cap landed 20 times upside down. His estimation of 20% showed he is using the results, but is not reasoning about percentages or proportions (e.g., 20 out of 300 is 6%). His level of reasoning is at a multi-structural level since he is making generalizations based on a large sample, and he is using the distribution of the outcomes to make his generalization, although not in the most appropriate way.

Variability. Greg reasoning about variability was not observed in Task4PreI. Greg recognized variability within and across samples in Task6Inst. He observed "which fish had more" when making his estimation of the probability to catch a Blue Bass. He observed Green Gills were caught the most in every sample he collected. His level of reasoning in the variability construct is multi-structural. In Task7Inst, Greg recognized variability within a single sample by noticing the little amount of 5s, but he also recognized variability across samples since he noticed that 5 came up very little after collecting 42, 561, and 100 trials. His level of reasoning during this task stayed at a multi-structural level. Greg suggested dropping the toothpaste cap 300 times "because there are three sides" and he mentioned that "using logical… it will land more in the side because it has more space to land on, but you never know." The interviewer asked if he thinks he will get 100 in the side, 100 upside down, and 100 right sided up out of 300 tosses to what he responded with a confident "no." His level of

reasoning is at a uni-structural level since he demonstrated to be expecting variability from the 300 times the toothpaste cap was drop.

Investigation. In the investigation construct, Greg reasoning was at a uni-structural level in Task4PreI since he propose to collect data, but did not mentioned how he will assess the results. He reasoned at a relational level in Task6Inst since he proposed an appropriate data collection process (several samples of size 30) and reasoned appropriately on how to assess the results (using the distribution of the outcomes). In Task7Inst Greg proposed an appropriate data collection process as well, collecting several samples of size 42, 561, and 100; and he used the distribution of the outcomes to assess the results. His level of reasoning is at a relational level. In Task11PostI Greg proposed an appropriate data collection (e.g., a sample of size 300) and proposed to use the distribution of the outcomes to assess the results. His level of reasoning is at a multi-structural level.

Table 6.4. Greg's SOLO level in each task – 4th sequence

	Task4PreI	**Task6Inst**	**Task7Inst**	**Task11PostI**
Generalization	Uni-structural	Multi-structural	Multi-structural	Multi-structural
Variability	Not observed	Multi-structural	Multi-structural	Uni-structural
Investigation	Uni-structural	Relational	Relational	Multi-structural

Greg increased his level of reasoning in all four sequences in the generalization and the investigation constructs. His level of reasoning stayed constant during the 1st and 2nd sequence in the variability construct, and after he demonstrated to be reasoning at a multi-

structural level he moved back to a uni-structural level of reasoning during the 3rd sequence. He also demonstrated to be reasoning at a relational level in the generalization and investigation constructs during the 3rd sequence. If the instructional tasks are not taken into account in the 4th sequence, it could be said that Greg increased his levels of reasoning in the generalization and investigation constructs.

The case of Jasyn

Jasyn's 1st sequence

This sequence includes three tasks where students have to reason about the fairness of a coin, a bottle cap, and a shaped button, respectively. Only two of the three construct were used to analyze changes in probabilistic reasoning in this sequence: generalization and variability.

Generalization. Jasyn reasoned at a uni-structural level when generalizing about the fairness of the coin, the bottle cap, and the shaped button. At first, using a sample of size 10, he generalized that the coin was unfair "because it [was] weighted", but after 100 and 1000 tosses he noticed the difference between the number of heads and tails generalizing the coin was unfair because there were more heads than tails. In Task8PostT, he observed the difference between the number of times the bottle cap landed up and down and he generalized the bottle cap was fair "because they are not always even." And in a similar way than in Task1PreT, after 100 and 1000 tosses he generalized the bottle cap was unfair because the number of times it landed up was "almost twice as many as down." His

reasoning is at a uni-structural level since, although he recognized the difference between the outcomes, he made generalizations after 10 tosses not recognizing the importance of sample size. His level of reasoning stayed at a uni-structural level in Task12RetT since he makes generalizations about whether it was more likely for the button to land with the ridged side up even after 10 tosses, not recognizing the importance of sample size to make generalizations. He gave the same reason after 10, 100, and 1000 tosses stating it was more likely the button landed with the ridged side up "because the ridged side weighted more."

Variability. When reasoning about variability, Jasyn recognized the variability among results within a single sample in Task1PreT and Task8PostT, located at a uni-structural level. He noticed "tails is more than 2 times as much" in Task1PreT and that up "almost twice as many down" in Task8PostT. In Task12RetT he is not recognizing variability both within and across sample, locating him at a pre-structural level. It seems like the context of the situation is driving his intuition during this task, Task12RetT. Recall the context of this task is a shaped button that has a ridged side. Since it seems to him the ridged side was heavier, he responded the button was not fair.

Table 6.5. Jasyn's SOLO level in each task – 1st sequence

	Task1PreT	**Task8PostT**	**Task12RetT**
Generalization	Uni-structural	Uni-structural	Uni-structural
Variability	Uni-structural	Uni-structural	Pre-structural

Jasyn's 2nd sequence

This sequence asked to describe what he will do to determine the fairness of a coin (Task2PreT), a button (Task9PostT), and a bottle cap (Task13RetT), and what would give him convincing evidence.

Generalization. Jasyn's level of reasoning increased from pre-structural in Task2PreT to uni-structural in Task9PostT and moved back to pre-structural in Task13RetT. In Task2PreT he did not make generalizations about the fairness of the coin stating he will "let them chose what side they wanted every time", reasoning at a pre-structural level. In Task9PostT he propose to use a sample of size 100 to make generalizations about the fairness of the bottle cap, but did not described how he will assess the results, reasoning at a uni-structural level. In Task13RetT he mentioned that "nothing would have to convince [him] the ridges side weights more" believing that all bottle caps that have a ridge side are unfair. Since he did not generalize about whether the button was more likely to land on the ridge or the flat side he is reasoning at a pre-structural level. Similar than in Task12RetT, it seems like the context of the situation is driving his intuition in Task13RetT.

Variability. Reasoning about variability was not observed in any of the tasks. Jasyn did not recognize variability either within a single sample or across samples, reasoning at a pre-structural level in all three tasks of the sequence.

Investigation. Previous to instruction, in Task2PreT, Jasyn did not propose a data collection process and his answer has nothing to do with his intuitions or previous

experiences with coins. His reasoning at the beginning of the sequence is located at a pre-structural level. Jasyn only proposed a data collection process in Task9PostT, although he did not describe how to assess the result obtained from the 100 tosses he proposed. In this task his reasoning was at a uni-structural level. By the end of the sequence, Task12RetT, he moved back to a pre-structural level of reasoning since he did not proposed a data collection process and state that "nothing would have to convince [him] the ridges side weights more." As mentioned earlier, the context of the task seems to be driving his intuitions.

Table 6.6. Jasyn's SOLO level in each task – 2nd sequence

	Task2PreT	**Task9PostT**	**Task13RetT**
Generalization	Pre-structural	Uni-structural	Pre-structural
Variability	Pre-structural	Pre-structural	Pre-structural
Investigation	Pre-structural	Uni-structural	Pre-structural

Jasyn's 3rd sequence

Three tasks were included in this sequence. Task3PreI asked students to determine the number of times they needed to draw a single marble, out of a bag containing 10 marbles, in order to be very confident about how many marbles of each color are in the bag. Task 10PostI was similar to Task3PreI but using a jar of Jolly Ranchers instead of a bag of marbles. Task 5Inst asked to determine the content of a bag of marbles just knowing that it contains 12 marbles. No information about the colors was given.

Generalization. Jasyn uses small samples to make generalizations about the marbles

in the bag in Task3PreI. He makes generalizations with a sample of size three in the first part of the task and a sample of size eight in the second part of the task. His reasoning is at a uni-structural level since his generalizations are based on small samples and he did not used the distribution of the results to make generalization about the content of the bag. In Task5Inst Jasyn reasoned at a multi-structural level in the generalization construct. Even though He based his best predictions in the outcomes as he stated "[he] got 4 the most."

After instruction, in Task10PostT, Jasyn made generalization about the content of the jar of Jolly Ranger using samples of size 100 and 200. When asked what does it tells him if after 100 tosses he got 83 cherries and 17 grapes he answered "7 cherries and 3 grapes." When asked why, he answered "because 83 plus 17 is a hundred and I just went down a little more because that might not be right. So I went down with that one [the cherries] and gave that one [grapes] a little bit more." In this task Jasyn's level of reasoning is at a relational since he makes generalizations based on samples of large size and used the distribution of the outcomes to make generalizations about the content of the jar.

Variability. Jasyn does not showed evidence he recognized variability in Task3PreI, locating him at a pre-structural level in the variability construct. In Task5Inst he noticed the variability across samples recognizing 4 appears the most in the outcomes. From his final answer of 4 red, 4 black, and 4 yellow, it does not appear he paid attention to the variability within samples when making generalizations, locating him at a uni-structural level. In Task10PostT he seems to be expecting some type of variability within a sample when he said

"[he] went down a little more because that might not be right" after getting 83 cherries and 17 grapes out 100 trials. He noticed that obtaining 83 cherries and 17 grapes does not gave him the true content of the jar, but an estimation about what could be inside the jar. Since he recognized variability within sample and not across sample his level of reasoning stayed at a uni-structural level.

Investigation. In Task3PreI he proposed data collection and does not proposed how to assess the results, reasoning at a pre-structural level. During instruction, Task5Inst, Jasyn proposed to collect several samples and proposed to look at the distribution of the outcomes to assess the result. He is reasoning at a relational level in the investigation construct. After instruction Jasyn still reasoning at a relational level, in Task10PostT, since \he proposed a data collection and also proposed how to assess the results using the distribution of the outcomes.

Table 6.7 Jasyn's SOLO level in each task – 3rd sequence

	Task3PreI	**Task5Inst**	**Task10PostI**
Generalization	Uni-structural	Multi-structural	Relational
Variability	Pre-structural	Uni-structural	Uni-structural
Investigation	Pre-structural	Relational	Relational

Jasyn's 4th sequence

This sequence is the longest including four tasks. Task4PreI had students reasoning about the likelihood of a plastic cup to land upside down. Task6Inst presented a situation

about a lake that has two types of fish in it, and they needed to estimate the probability to catch a Blue Bass. In Task7Inst students needed to determine whether or not the dice sold by a given company was fair or not, and Task11PostI was similar to Task4PreI but using a toothpaste cap instead of a plastic cup.

Generalization. In Task4PreI Jasyn is reasoning at a pre-structural level since there is lack of evidence he wanted to use the distribution of the outcomes to make generalizations and she seems not to recognize the sample could tell her something about the underlying population. During Task6Inst, he had to determine the probability that a fish caught in a lake was a Blue Bass. He makes generalizations about the probability based on six samples of size 30 each. In order to estimates the probability that a fish caught is a Blue Bass or a Green Gill he uses the distribution of the outcomes, calculating the average and converting them to a percent. His estimation of 70% probability to catch a Green Gill and 30% probability to catch a Blue Bass was pretty close to the true underlying population of 1 Blue Bass per 2 Green Gills. His level of reasoning in this task is a multi-structural since he is not considering sample size.

In Task7Inst, the Schoolopoly task, Jasyn investigated the fairness of the dice sold by Slice -N- Dice with his partner Greg. Jasyn (same as Greg) recommended the school "should probably not buy a dice from this company because there is barely any five." The details about their analysis were presented above in Greg's analysis. His level during this task is multi-structural since he makes generalizations using the counts, instead of percentages or

proportions, on the distribution of the outcomes from several samples of appropriate sizes. In Task11PostI his generalization about the likelihood of the toothpaste cap to land upside down were based on a single sample of size 15 and he did not use the distribution of the outcomes to make generalizations, reasoning at a uni-structural level.

Variability. Jasyn does not recognize variability in Task4PreI since he is expecting to get all three outcomes with a very small sample (n=4). His reasoning here is at a pre-structural level. Since Task6Inst and Task7Inst consisted of a worksheet and a poster Greg and Jasyn completed together, Jasyn's analysis in both tasks are the same as Greg's. Their reasoning during Task6Inst and Task7Inst was at a multi-structural level, as they recognized variability within and across samples. For more details about their reasoning refer back to Greg's variability analysis during the 4th sequence in the previous section. Jasyn's level of reasoning is at a pre-structural level in Task11PostI since again he is not recognizing variability in small samples expecting to get all three possible outcomes from a sample of size 15.

Investigation. Jasyn proposed to collect data, but did not mentioned how he would assess the results, reasoning at a uni-structural level in Task4PreI. He reasoned at a relational level in Task6Inst since he proposed an appropriate data collection process (several samples of size 30) and reasoned appropriately on how to assess the results (using the distribution of the outcomes). In Task7Inst Jasyn proposed an appropriate data collection process as well and used the distribution of the outcomes to assess the results. His level of reasoning here is

relational. Jasyn proposed a data collection, although an inappropriate one (e.g., a sample of size 15), and does not propose to use the distribution of the outcomes to assess the results in Task11PostI. His level of reasoning is at a uni-structural level.

Table 6.8. Jasyn's SOLO level in each task – 4th sequence

	Task4PreI	**Task6Inst**	**Task7Inst**	**Task11PostI**
Generalization	Pre-structural	Multi-structural	Multi-structural	Uni-structural
Variability	Pre-structural	Multi-structural	Multi-structural	Pre-structural
Investigation	Uni-structural	Relational	Relational	Uni-structural

Jasyn did not show an increase on his level of reasoning during the 1st and the 2nd sequence, reasoning at a pre-structural and uni-structural level. It is surprising thought that he demonstrated to be reasoning at a relational level by the end of the 3rd sequence in the generalization and investigation constructs. During the 4th sequence, even if the instructional tasks are not taken into consideration, Jasyn's levels of reasoning in the variability and investigation constructs remained the same from the beginning to the end of the sequence. He only seems to increase his level of reasoning in the generalization construct, from Task4PreI to Task11PostI.

In summary, Greg and Jasyn demonstrated to increase their level of reasoning during the instruction. During the 1st and the 2nd sequence they mostly reasoned at a pre-structural and uni-structural level. Although Greg showed to be reasoning at a multi-structural level by

the end of the 1st and 2nd sequence in the generalization construct, Jasyn's level of reasoning seems to be inconsistent between pre-structural and uni-structural. During the 3rd and 4th sequence, both, Greg's and Jasyn's level of reasoning either remained the same or increase when the instructional tasks (Task5Inst, Task6Inst, and Task7Inst) are removed. Both, Greg and Jasyn, showed evidence to be reasoning at a relational level by the end of the 3rd sequence in the generalization and investigation construct.

Table 6. 9 Greg's and Jasyn's SOLO levels in all four sequences

Construct	Sequence	1st			2nd			3rd			4th			
	Task	Task1PreT	Task8PostT	Task12RefT	Task2PreT	Task9PostT	Task13RetT	Task3PreI	Task5Inst	Task10PostI	Task4PreI	Task6Inst	Task7Inst	Task11PostI
G	Greg	U	U	M	P	M	M	M	M	R	U	M	M	M
	Jasyn	U	U	U	**P**	**U**	**P**	U	M	R	P	M	M	U
V	Greg	U	U	U	P	P	P	**U**	**M**	**U**	N/O	M	M	U
	Jasyn	**U**	**U**	**P**	P	P	P	P	U	U	P	M	M	P
I	Greg				P	U	U	U	M	R	U	R	R	M
	Jasyn				**P**	**U**	**P**	P	R	R	U	R	R	U

CHAPTER 7

CROSS-ANALYSIS

This chapter presents a comparison of the probabilistic inference of the six case-study students in general. Tables are included in order to gather the information across students within each sequence, as well as within students in the order the data was collected. The later can help consider how things changed overall as time progressed across data collection points - before instruction, during instruction, after instruction. This view also allows one to consider the mode in which the student was engaging in the task (in a one-on-one interview, a written test, or working with a partner using software over an extended period of time).

This chapter is divided into two main sections. The first includes a deeper analysis across the students' performance at each construct within each sequence. The second analyzes performance of each student individually in each construct across the tasks in the order they were collected before, during, and after instruction.

Across students within each sequence

This section looks at the data across all six students, comparing their reasoning within each sequence. It consists of three main sub-sections (generalization, variability, and investigation) where students' level of reasoning are compared using each of the four sequences.

Generalization

The generalization construct was used to observe the way students were making generalizations beyond a data collected. Some students used their intuitions and ignored the fact that sampling could provide a good estimation about the underlying population distribution. Others reasoned and made generalizations based on small samples using the strategies mentioned by Lee (2005) such as the "total weigh approach." And others, although not reasoning appropriately about how to use the result to make generalizations about the underlying population, did propose a large sample size. Table 7.1 illustrates the SOLO levels used to characterize each student after analyzing their answers to each of the tasks included in the four sequences. Bolded borders separate tasks for each sequence. The SOLO levels are represented with the letters P (pre-structural), U (uni-structural), M (multi-structural), and R (relational).

Within the 1st sequence most of the students showed some progression, moving up at least one level in their reasoning, while some of them stayed at the same level of reasoning. Only one out of the six students (Manuel) seemed to use less sophisticated reasoning after instruction, in Task12RetT.

During this sequence the students were given distributions of outcomes after 10, 100, and 1000 tosses of a coin, a button, and a bottle cap, respectively. Lara, Dannie, Brandon, and Greg appeared to recognize the importance of sample size to make generalizations, acknowledging it is not appropriate to use a sample of size 10 to make generalizations about

the likelihood of the button to land with the ridged side up by the end of the sequence. Even more, although Lara used a sample of size 100 to make generalizations, Brandon wasn't sure and trusted more in the results from a sample of 1000 tosses. Dannie was the only one who reasoned connecting both samples (n=100 and n=1000) to make generalizations stating the button was unfair "because of what happened the other times" noticing the distribution of the outcomes after 10 tosses and after 100 tosses as well. Manuel and Jasyn do not seem to acknowledge the importance of sample size within the task in this sequence as their reasoning stayed for the most part at the uni-structural level.

Table 7.1. *Generalization*: SOLO Levels for each students per task per sequence

Sequence	1st			2nd			3rd			4th			
Task	Task1PreT	Task8PostT	Task12RetT	Task2PreT	Task9PostT	Task13RetT	Task3PreI	Task5Inst	Task10PostI	Task4PreI	Task6Inst	Task7Inst	Task11PostI
Lara	P	M	M	P	M	M	P	U	M	U	M	M	M
Dannie	U	M	R	P	M	M	U	M	M	P	M	M	M
Manuel	U	P	P	U	U	M	P	M	M	U	M	R	U
Brandon	P	P	R	P	P	U	U	M	M	P	M	R	U
Greg	U	U	M	P	M	M	M	M	R	U	M	M	M
Jasyn	U	U	U	P	U	P	U	M	R	P	M	M	U

The context in the 2nd sequence is the same as the 1st sequence with the main difference that they were given no data, but a situation where they were asked what they will

do and what will give them convincing evidence about the fairness or the likelihood for an event to happen. During this sequence, all students but Brandon and Jasyn recognized a large sample was needed in order to make generalizations. Brandon only used a small sample to make generalizations in Task13RetT. Jasyn, after using a small sample in Task9PostT, used the physical appearance of the bottle cap (Task13RetT) and decided the ridged side was more likely because it weighted more, thus lowering his reasoning level. Also, none of them reasoned at a relational level in any of the tasks in the 3rd sequence, which implies they did not use several samples of appropriate size nor explicitly used the distribution of the outcomes to make generalizations.

During 3rd sequence all six students showed an increase in their reasoning levels concerning generalization. By the end of the sequence all students made generalizations based on the results of a single sample of an appropriate size (e.g., n=100) or several samples of small size (e.g., n=12). During the instructional task five of them used the "total weight approach" to decide on the number of trials. They knew the bag consisted of 12 marbles so they used mostly eight samples of size 12, with the exception of Jasyn who used eight samples of size eight. During that task they did not know how many colors were in the bag, so it is not clear were Jasyn's decision of eight trials came from.

The 4th sequence included four tasks, where two of them where completed in pairs and two on a one-on-one interview. Paying attention to students' levels in the interview tasks, Task4PreI and Task10PostI, all students either increased or stayed the same in their level of

reasoning. Lara and Dannie increased the most, moving from the lower levels to the upper levels of reasoning. Manuel and Jasyn, who stayed at the same level of reasoning, only used small samples and did not use the distribution of the outcomes to make generalizations about the content of the bags of marbles and the jar of Jolly Ranchers. All students reasoned at a multi-structural or relational level during the instructional tasks (Task6Inst and Task7Inst).

Variability

The variability construct was the hardest to characterize. This construct seeks for evidence of students' awareness of variability within and between samples, and their tolerance of variability. This construct was characterized using all tasks during the four sequences. Table 7.2 illustrates the SOLO levels used to characterize each student after analyzing their answers to each of the task included in the four sequences. The SOLO levels are represented with the letters P (pre-structural), U (uni-structural), M (multi-structural), and R (relational). N/O means not observed.

During the 1st sequence four out of six of the students reasoned either at a pre-structural or uni-structural level, recognizing no variability at all or recognizing variability only within a single sample. Only Dannie showed evidence of recognizing variability within and between samples by the end of the sequence (multi-structural), recognizing only variability within a single samples in the first two taks (uni-structural). Greg recognized variability within a single sample from the beginning of the sequence to the end (uni-structural).

In the 2nd sequence most of the students (4 out of 6) reasoned at a pre-structural level from beginning to end of the sequence showing no recognition of variability at all. Lara and Dannie (the high achieving pair) showed some awareness of variability within samples by the second task, but only Dannie increased her level to multi-structural recognizing variability among and between samples.

Table 7.2. *Variability*: SOLO Levels for each students per task per sequence.

Sequence	1st			2nd			3rd			4th			
Task	Task1PreT	Task8PostT	Task12RetT	Task2PreT	Task9PostT	Task13RetT	Task3PreI	Task5Inst	Task10PostI	Task4PreI	Task6Inst	Task7Inst	Task11PostI
Lara	P	U	U	P	U	U	P	U	U	P	M	M	U
Dannie	U	U	M	P	U	M	P	M	P	N/O	M	M	U
Manuel	U	P	P	P	P	P	P	U	U	N/O	M	M	P
Brandon	P	P	U	P	P	P	U	U	P	N/O	M	M	N/O
Greg	U	U	U	P	P	P	U	M	U	N/O	M	M	U
Jasyn	U	U	P	P	P	P	P	U	U	P	M	M	P

The 3rd sequence shows the same pattern where almost all of them reasoned at a uni-structural level. In this sequence only Greg showed he recognized variability within and across samples. Lara, Manuel, and Jasyn showed no awareness of variability at the beginning of the sequences and then increased their reasoning to a uni-structural level, recognizing variability within a single sample. Although Dannie showed awareness of variability during

the instruction, he seems not to be aware of it during the post task, Task10PostI, reasoning at a pre-structural level.

During the 4th sequence students were not prompted to provide reasons for their responses too much by the interviewer in Task4PreI and Task11PostI, so in some occasions reasoning about variability was not observed. In Task6Inst and Task7Inst the six students reasoned at a multi-structural level, meaning they recognized variability within and between samples. In Task11PostI reasoning about variability was not observed for Brandon, three of them (Lara, Dannie, and Greg) recognized variability within a single sample, and Manuel and Jasyn did not recognize variability.

Investigation

This construct refers to the ability students have to propose an appropriate investigation process, including describing how the data will be collected and how the results will be analyzed. It was not appropriate to characterize students' investigation construct during the 1st sequence since samples were already collected, and the distribution of the outcomes were part of the task for them to analyze. The context in the 2nd sequence is the same as the 1st sequence with the main difference that in this one they were given no data, but a situation where they were asked what they will do and what will give them convincing evidence about the fairness or the likelihood for an event to happen. Table 7.3 illustrates the SOLO levels used to characterize each student after analyzing their answers to each of the task included in the four sequences. The SOLO levels are represented with the letters P (pre-

structural), U (uni-structural), M (multi-structural), and R (relational).

Students' reasoning within the investigation construct varies from thinking they were not able to estimate the content of a bag of marbles unless they looked at it (pre-structural), to proposing to collect data, considering sample size, and describing how the data collected should be assessed in order to estimate the content of the bag (relational). Some students started proposing small samples and increased their reasoning considering sample size as an important piece on the data collection process. Others, although proposed a collection of data using a large sample size, did not utilize the distribution of the outcomes in the most appropriate ways.

Since the students were not asked to collect data or describe how they will assess the results obtained from the data collected, the 1st sequence does not include an analysis of the investigation construct.

Through the 2nd sequence, most of the students reasoned at the two lowest levels in the SOLO levels. Jasyn proposed a data collection process only in Task9PostT, Greg proposed to collect data on Task9PostT and Task13RetT, and Brandon proposed it only in the last task of the sequence (Task13RetT). Manuel stayed at a uni-structural level, only proposing to collect data but without describing how the result would be analyze. Dannie did not propose a data collection in the first task and moved all the way up to the highest level of reasoning during Task9PostT, proposing a data collection, recognizing the importance of sample size and describing how the data could be assess using the distribution of the

outcomes. During the last task, Task13RetT, he moved back one level still proposing a data collection recognizing sample size, but did not use the distribution of the outcomes appropriately to assess the results.

Table 7.3. *Investigation*: SOLO Levels for each students per task per sequence.

Sequence	1st			2nd			3rd			4th			
Task	Task1PreT	Task8PostT	Task12RetT	Task2PreT	Task9PostT	Task13RetT	Task3PreI	Task5Inst	Task10PostI	Task4PreI	Task6Inst	Task7Inst	Task11PostI
Lara				M	M	U	P	M	R	U	M	M	M
Dannie				P	R	M	U	M	M	U	M	M	M
Manuel				U	U	U	P	M	R	P	M	R	M
Brandon				P	P	U	U	M	U	P	M	R	U
Greg				P	U	U	U	M	R	U	R	R	M
Jasyn				P	U	P	P	R	R	U	R	R	U

All students but Brandon showed a sustained increase in their reasoning during the 3rd sequence. At the beginning, half of the students did not propose to collect data and the other half proposed to collect a small sample, but without mentioning how they will assess the results. Jasyn's level of reasoning increased from pre-structural (Task3PreI) to relational (Task5Inst), which mean he went from not proposing to collect data to proposing data collection of an appropriate size and reasoning appropriately about how to use the distribution of the outcomes to assess the results from the data collected. Lara and Manuel

also started reasoning at a pre-structural level during the pre-interview and increased their reasoning to the higher level by the time they were post-interviewed.

Throughout the 4th sequence all students showed an increase in their investigation reasoning. Lara, Dannie, Greg, and Jasyn started proposing to collect small samples, and Manuel and Brandon did not propose any data collection at all. Lara's and Dannie's levels of reasoning through this sequence progressed evenly. They moved to the next level in Task6Inst proposing to collect data but without recognizing the importance of sample size. They used the distribution of the outcomes appropriately to assess their results. In Task7Inst, the last task during the instruction, all but Lara and Dannie reasoned at a relational level. They all recognized the importance of sample size, collecting samples of large sizes, and using the distribution of the outcomes to determine whether or not the dice were fair. By the end of the sequence all but Brandon and Jasyn were reasoning at a multi-structural level. In the 4th sequence, paying attention to students' level of reasoning in the interview tasks, Task4PreI and Taks10PostI, all students but Jasyn increased their level of reasoning. Jasyn's level of reasoning stayed in the uni-structural level.

In summary, the four sequences showed an increment in most of the students' reasoning in the generalization construct meaning they are more aware about the importance of sample size to make generalization. Although some of the students did not use the distribution of the outcomes in the most appropriate way, they did use it in some way, recognizing that sampling could be used to gather information that could help them to make

inferences about the underlying population. It was also found that for the most part students reasoned at a pre-structural or a uni-structural level in the first three sequences. Within this construct it was found student reason at a higher level during the instructional tasks and the interviews during the 3rd and the 4th sequence.

Across students across tasks in the order collected

This section contains the analysis for students' performance across the tasks in the order they were collected. The main foci of the section are to: (1) examine whether or not the students' probabilistic inference reasoning increased, stayed the same, or decreased across the instruction, and (2) inspect if the mode in which the student was engaging in the task (written test, interview, paired instruction) seemed to influenced their ability to reason and in what way. It is also important to notice that the students had access to software during the instruction while working with their partners, but had no access to any type of technology during the written test and during the one-on-one interviews.

Changes Across Instruction

Generalization

In general, observing each row on Table 7.4, Lara, Dannie, and Greg showed to increase their level of reasoning the most, following a similar pattern (except for Greg's response to Task8PostT). Manuel and Brandon moved back and forth between the levels of reasoning, especially during the post-interview and the retention test. Jasyn generally exhibited lower levels (pre-structural and uni-structural), with the exception of the

instructional tasks.

Their reasoning levels in the tasks collected before instruction was mostly either pre-structural or uni-structural, and all of them reasoned either at a multi-structural or relational level during the instructional tasks. The level of questioning by the interviewer in Task10PostI seemed to encourage them to reason at a higher level. The interviewer did not question the students as much during Task11PostI, resulting in shorts, and sometimes vague, answers making it difficult to characterize their level of reasoning.

Table 7.4. *Generalization*: SOLO Levels for each students through instructional unit.

Task	Task1PreT	Task2PreT	Task3PreI	Task4PreI	Task5Inst	Task6Inst	Task7Inst	Task8PostT	Task9PostT	Task10PostI	Task11PostI	Task12RetT	Task13RetT
Lara	P	P	P	P	U	M	M	M	M	M	M	M	M
Dannie	U	P	U	P	M	M	M	M	M	M	M	R	M
Manuel	U	U	P	U	M	M	R	P	U	M	U	P	M
Brandon	P	P	U	P	M	M	R	P	P	M	U	R	U
Greg	U	P	M	U	M	M	M	U	M	R	M	M	M
Jasyn	U	P	U	P	M	M	M	U	U	R	U	U	P

During all the *pre* tasks (Task1PreT, Task2PreT, Task3PreI, and Task4PreI) Lara reasoned at a pre-structural level, she was not paying attention to sample size and was not using the distribution of the outcomes to make generalizations. In all instructional tasks (Task5Inst through Task7Inst) her level of reasoning increased to a multi-structural level,

being aware of the importance of sample size and using the distribution of the outcomes to make generalization, although not in the most appropriate way. She continued on a constant level of reasoning toward the end of the unit. Similar behavior is observed in Dannie's generalization reasoning. She started either making no generalizations or using small samples to infer characteristics from an underlying population, and during and after instruction she reasoned at a multi-structural level, using a larger sample to make a generalization and using the distribution of the outcomes in some way to assess the results. Dannie's reasoning was observed to be in the highest level (relational) during Task12Ret where she used several aspects of the situation, making connections about the different samples. In general, Lara and Dannie do not seem to be either beneficiated nor affected by the mode of engagement demonstrating a solid increase in their level of reasoning regarding the generalization construct. Lara and Dannie were the high-scoring pair selected at the beginning of the study.

Manuel and Brandon, the average-scoring pair, reasoned at either a pre-structural and uni-structural level during the *pre* tasks (Task1PreT, Task2PreT, Task3PreI, and Task4PreI) and showed an increase in their level of reasoning moving to the multi-structural and relational levels during the instructional tasks (Task5Inst through Task7Inst). This pair demonstrated wide variability in their levels of reasoning on different post tasks, albeit many times they were reasoning at almost the same level (Task8PostT through Task13RetT) showing the greatest difference in Task12RetT. With Manuel and Brandon the mode of engagement in the task enhanced their reasoning, having a steady performance during the instructional tasks where they worked together as a pair and had access to the software. Their

inconsistencies in levels of reasoning on the post tasks indicate that they were capable of reasoning at higher levels (M and R) on some tasks, but that perhaps they did not see the need to apply their understanding of generalization in all tasks.

Jasyn's level of reasoning demonstrated to be uni-structural, with only two pre-structural instances (Task2PreT and Task13RetT) and one relational instance (Task10PostI). The questioning within this particular task in the post-interview encouraged him to reason at a higher level. His reasoning demonstrated to be stable and higher during the instructional tasks. Greg level of reasoning increased from the *pre* tasks (Task1PreT through Task4PreI) to the instructional tasks (Task5Inst through Task7Inst). His level of reasoning was mostly uni-structural during the *pre* tasks (with the exception of Task3PreI) and multi-structural during and after instruction (with the exception of Task8PostT and Task10PostI).

Variability

In general, all students reasoned either at a pre-structural or a multi-structural level through the tasks before and after instruction, and either uni-structural or multi-structural during the instructional tasks. None of them reasoned at a relational level in this construct. And in some cases their variability reasoning was not observed, indicated with an N/O in Table 7.5.

Lara was the only student that showed an increase in her level of reasoning, starting on the pre-structural level during the *pre* tasks (Task1PreT through Task4PreI) and moving to uni-structural during all *post* tasks (Task8PostT through Task13RetT). Her reasoning was higher

during instruction reasoning at multi-structural level. Dannie's level of reasoning was mostly pre-structural previous to instruction, showing an increased during the instructional tasks (multi-structural). In Task5Inst, even though they worked in pairs they reasoned individually since each one had to "guess" the content of a bag his/her partner designed. Dannie showed an increase in her reasoning after instruction moving up to a uni-structural level during the post-test and the post-interview (except Task10PostI) and even more reasoning at a multi-structural level during the retention test.

Table 7.5. *Variability*: SOLO Levels for each students through instructional unit.

Task	Task1PreT	Task2PreT	Task3PreI	Task4PreI	Task5Inst	Task6Inst	Task7Inst	Task8PostT	Task9PostT	Task10PostI	Task11PostI	Task12RetT	Task13RetT
Lara	P	P	P	P	U	M	M	U	U	U	U	U	U
Dannie	U	P	P	N/O	M	M	M	U	U	P	U	M	M
Manuel	U	P	P	N/O	U	M	M	P	P	U	P	P	P
Brandon	P	P	U	N/O	U	M	M	P	P	P	N/O	U	P
Greg	U	P	U	N/O	M	M	M	U	P	U	U	U	P
Jasyn	U	P	P	P	U	M	M	U	P	U	P	P	P

Manuel, Brandon, and Jasyn reasoned either at a pre-structural or a uni-structural level previous to instruction and after instruction. In general, they did demonstrate to improve their reasoning as most of their instances were at a pre-structural level. In the other hand, Greg seemed to be reasoning mostly at a uni-structural level with only a few instances

of a pre-structural level. He also reasoned at a multi-structural level when "guessing" the content of the bag of marble Jasyn designed (Task5Inst).

In this construct the mode of engaging in the tasks seems to influence students' reasoning. As it is observed in Table 7.5, five of the students reasoned at a higher level when working in pairs. Also the way the worksheets were designed could help in enhancing students' reasoning. The directed questioning and the inclusion of tables for them to fill out might encourage them to recognize the differences among the individual results within each sample as well as to recognize the different outcomes obtained from different samples.

Investigation

This construct showed the greater increase across the students, again if analyzing their performance during instruction separately. It is observed from Table 7.6 that most of the instances in the tasks completed previous to instruction are either pre-structural or uni-structural. Observing the instances in the task after instruction, there are several multi-structural reasoning and few relational reasoning, the two higher levels of reasoning in the probabilistic interference framework. During the instructional tasks all students reasoned either at a multi-structural or relational level.

Lara, Dannie, and Manuel showed an increment on their level of reasoning, although it is observed some uni-structural instances in Lara's Task13RetT and Manuel's Task9PostI and Task13Ret. Five out of the six students reasoned at a relational level at some point after the instruction (with the exception of Brandon).

Table 7.6. *Investigation*: SOLO Levels for each students through instructional unit.

Task	Task1PreT	Task2PreT	Task3PreI	Task4PreI	Task5Inst	Task6Inst	Task7Inst	Task8PostT	Task9PostT	Task10PostI	Task11PostI	Task12RetT	Task13RetT
Lara		M	P	U	M	M	M		M	R	M		U
Dannie		P	U	U	M	M	M		R	M	M		M
Manuel		U	P	P	M	M	R		U	R	M		U
Brandon		P	U	P	M	M	R		P	U	U		U
Greg		P	U	U	M	R	R		U	R	M		U
Jasyn		P	P	U	R	R	R		U	R	U		P

From the written test it is hard to tell whether students' reasoning increased, stayed the same, or decreased given that one of the sequences (1st sequence) was not used to characterize student's probabilistic inference on the investigation construct. The interviews seemed to collect more evidence about students' reasoning about a data collection process and the use of the distribution of outcomes to make sense of the data collected. All students but Brandon and Jasyn increased their level of reasoning from the pre-interview to the post-interview. Again the instructional task enhanced students' thinking and reasoning about the investigation process, since students were encouraged to collect their own data and use it to make estimations and predictions about an underlying population.

Impact of Mode on Levels of Reasoning

Written tests. From the pre-test, post-test, and the retention test it was not observed significant changes in students' reasoning level. The most significant differences were observed within the generalization construct (Table 7.4) since they increased the number of their samples, and based their generalizations in the distribution of the outcomes. From Tables 7.4 through 7.6 it is more frequently observed pre-structural and uni-structural instances, mostly in the variability and investigation construct. Characterizing students' level of reasoning in the variability construct using a written test was a challenge because their arguments were usually short and vague. It was easier to use written test to characterize students' reasoning about generalization, where the bigger changes were observed.

Interviews. Sometimes interviews make students reason at a higher level, depending on the questioning of the interviewer. With interviews, students are more open to talk about their ideas and be challenged with additional questions and arguments. In the pre-interview the students showed to be reasoning at a pre-structural or uni-structural level across the three constructs. By the time they were post-interviewed they showed an increase in their reasoning, mostly in the generalization and investigation construct. Interviews seem to be beneficial to unveiling students' reasoning. The inquiry by the interviewer may give them more opportunities to reason at higher levels, when they have been taught the material and have arguments to respond to those inquiries.

Paired work on instructional tasks. As it is observed in Tables 7.4-7.6 all students

reasoned at a multi-structural or relational level during the instructional tasks where they worked in pairs, Task5Inst, Task6Inst and Task7Inst. During these tasks they were working in pairs using software to conduct their simulations. They worked in pairs for about 12 days, negotiating about the results they got, and making decisions based on those negotiations. The use of the software could be a motivation to run multiple samples of large sizes, since it is easy and quick. Also the use of worksheets with explicit instructions to follow might have also influenced their reasoning and impel it to the higher levels. The worksheets included questions they needed to answer and tables they had to fill out, and in some occasions the tables included the number of trails to run (see Appendix A). The context of the task could also have encouraged higher levels of reasoning since they were more engaged in the task and were motivated to think and find an answer to the problem, as in the case of the Schoolopoly Task (Task7Inst).

In summary, when trying to characterize students' level of reasoning, the interviews and the work with partners seem to be the most helpful mode of engagement for students. The negotiation that is assumed to occur while they work in pairs seems to be beneficial to enhance students' reasoning levels to higher levels. Interviews are a great tool to assess students about previous and current knowledge. From the interviews, the researcher was able to investigate whether or not each student's probabilistic inference increase, stayed the same, or decreased.

CHAPTER 8

DISCUSSION

The purpose of this study is to investigate how middle school students' ability to engage in probabilistic inference changes over the course of instruction. A Probabilistic Inference Framework (Table 2.2) has been developed in order to characterize student's learning of informal probabilistic inference. Recall that probabilistic inference is equivalent to statistical inference in a probabilistic context.

The framework is based on existent work on students' probabilistic thinking (Jones et al., 1997), informal statistical inference (Makar & Rubin, 2009; Zieffler et al., 2008), and students' inferential statistical inference (Watson, 2001; Vallecillos & Moreno, 2002). Although there are several frameworks in the literature (Jones et al., 1997; Makar & Rubin, 2009; Vallecillos & Moreno, 2002), there is not a framework to study students' inferential statistical inference in a probabilistic context. Three constructs were identified to be important key principles to be included in the framework: *generalization, variability, and investigation*. For that reason the author drew on literature to develop the framework explained in Chapter 2 and used it to analyze the data.

Researchers have called for more research in the area of probabilistic inference in order to gain knowledge about how students learn statistics, especially how they use data to infer characteristics of the underlying population (Jones et al., 2007). They have also called for more teaching experiments in probability in order to determine the impact instruction has

on students' informal statistical inference. This investigation responded to that call and aims to contribute to the field with additional results and knowledge about the effect instruction has in students' probabilistic inference.

This chapter is divided into four sections: summary and conclusions, limitations, implications, and recommendations. A summary of the study is presented, followed by conclusions of the research questions. Next, the limitations will be described. A discussion of implications for teachers will follow. Finally, recommendations for future research will be addressed.

Summary and Conclusions

This study utilizes case study research. Six students were selected based on their scores on a standardized mathematics achievement test and a pre-test on probability concepts. Students were engaged in a series of challenging tasks in a 12-day instructional program in probability. Six problem-based tasks were used during the 12 days of instruction where students were required to use a computer-based simulator to collect, display, and analyze data. A sequence of tasks focused on similar constructs was analyzed in order to study how students' abilities to engage in probabilistic inference changes. This study intends to bring some insights on how students reason about important concepts in informal statistical inference using probabilistic situations, and how that reasoning is influenced by instruction. The author paid special attention to students' abilities to make probabilistic inferences taking into account the role of variability and sample size, their ability to make

generalizations, and, when possible, their abilities to propose an investigation process. The data was used to answer the following research questions:

1. How do middle school students' abilities to make probabilistic inferences change over the course of instruction?
2. How do middle school students' perception about the importance of sample size when making probabilistic inference change?

Research Question #1

Growth in the ability to make probabilistic inference means for the student to be able to make generalizations beyond the data using the empirical distribution to make prediction about the theoretical probability, recognizing the importance of sample size and types of variability, and, when possible, their abilities to propose an investigation process.

All tasks included in the analysis asked students to make generalizations beyond the data on hand; from predicting the distribution of a bag of marble (knowing the total of marbles as well as how many different colors were in the bag) to estimating the probability distribution of a die sold by certain company. Some students used their intuitions and ignored the fact that sampling could provide a good estimation about the underlying population distribution. Others reasoned and made generalizations based on small samples using the strategies mentioned by Lee (2005) such as the "total weight approach." And others, although not always reasoning appropriately about how to use the result to make generalizations about the underlying population, did propose a large sample size.

As it is shown in Table 7.4, five out of six students demonstrated a growth in making generalizations beyond data, seemed to recognize the importance of sample size during and after instruction, although not all of them showed consistency in their reasoning levels during the tasks after instruction. Researchers have found the use of simulation tools help students develop a better understanding of empirical probability as it relates to sample size (Aspinwall & Tarr, 2001; Lee et al., 2009; Stohl & Tarr, 2002a). Similar to results found during this study, Aspinwall and Tarr (2001) found the use of technology foster growth in students' probabilistic reasoning. Results reported in Lee et al. (2009) are a deep analysis of the six case-study works during the Schoolopoly task. Their results also indicate that after appropriate instruction students recognize the importance of variability as well as the importance of sample size to make generalization beyond data.

Students' ability to recognize variability was hard to measure in some of the modes in which the students were engaging in the task, such as the one-on-one interviews. Students' levels of reasoning either increased or stayed the same, but there was not enough information to definitively say there was no increased in their awareness of variability. Their awareness seemed to be higher during the instructional tasks given they were working in pairs with access to a simulator tool. As researchers have found the understanding of variability seems to be related with the context of the task (Rubin et al., 2006) or to the design of the instruction (Ben-Zvi, 2006). The results from this study add that recognizing variability seems to be enhanced depending on the mode of engagements of the tasks (e.g., interviews, and instructional tasks). It was also found that the format of the worksheet used by the

students enhanced their recognition of variability within and across samples. As it was discussed in Chapter 7, the higher level of reasoning regarding the variability construct was found during the instructional tasks (see Table 7.5). In these tasks students were challenged to solve a problem that was meaningful to them, as recommended by Zieffler et al. (2008).

Throughout the tasks it was demonstrated students incorporate expressions about uncertainty, such as "there were probably more greens", "each one has a pretty good chance", and "it will land more in the side… but you never know." Paparistodemou and Meletiou-Mavrotheris (2008) also found that students can incorporate expressions about uncertainty. The incorporation of appropriate probabilistic language is a key aspect in students' informal inference, as indicated by Makar and Rubin (2009).

Most tasks included in this study required the collection of data, where students needed to propose a sample size, and to describe how they will make sense of the data collected. Students seemed to be aware about the importance of data collection, but did not always offer information on how to assess the results before instruction; however, they showed an increase in this ability by the end of the instruction, being able to propose a data collection method and make reference to the distribution of the outcomes to assess their results. As it is shown in Table 7.6, all students increased their level of reasoning during the instructional tasks. After the instructional tasks, their level of reasoning showed to be applied inconsistent between different modes of instruction/engagement (higher within interviews, more inconsistency within written tests).

In summary, the written tests used in this study are not the best tool to analyze students' changes in probabilistic inference. Although interviews had a protocol to follow, interviewers did not prompt students for explanation in the same manner. It is necessary to have a more explicit protocol with specific questioning intended to gather evidence of students' reasoning. The most appropriate tasks for promoting growth in probabilistic inference during this study were the instructional tasks, where students had access to a computer-based simulator and worked with a partner.

As indicated by Zieffler et al. (2008) the type of task selected when conducting research in order to study students' informal statistical inference is of great importance. Such tasks must provide students with opportunities to reason about a situation, make conjectures, and use data to support these conjectures. Interviews in the study fail in that part since students were presented with situations, but were not given opportunity to test their conjectures and support them with data. On the other hand, they had those opportunities during the instructional tasks. Zieffler et al. also mentioned the necessity of using challenging tasks to enhance students' understanding about informal statistical inference, a crucial characteristic that the Mystery Fish in a Lake Task and the Schoolopoly Task shared.

Research Question #2

As indicated by Pfannkuch (2005), recognizing the importance of sampling and sample size is one to the biggest problems related with informal inference and with it the interconnection between sample size and variability. Sample size is one of the main ideas in

statistics and sometimes it is hard for students to understand its importance. For most of the time the six case-study students believed that either small or large samples were representative of the underlying population. Even during the instructional tasks, they often used several samples of small sizes to predict the distribution of the colors in the bag of marble, and to estimate the proportions of Blue Bass fish in the Mystery Fish on a Lake task.

NCTM (2000) recommends students in the secondary levels carry out simulations of random phenomena, to interpret results obtained from these simulations, and to make inferences about the underlying population based on data. Connected to the idea of making inferences about an underlying population is the importance of sample size. In order to draw inferences a sample has to be representative of the underlying population. In order for that to happen it is necessary for the sample size to be large. This will reduce variability and the distribution of the outcomes will be more representative of the underlying population that those obtained with small samples.

Through the study, it was observed students' uses of small samples (n=4) to large samples (n=5000). During the *pre* tasks (Task1PreT through Task4PreI) most of the students were not aware of the importance of sample size in order to draw inferences about the underlying population. Only one was able to recognize it was not appropriate to use a sample of size 10 to make inferences about the underlying population. Five out of six students only used the "total weight approach" discussed by Lee (2005) during the *pre* tasks.

A change of students' perception about sample size was noticed when the students

were engaged in the instructional tasks. Although they still were using small samples, following the "total weight approach", they were using more than a single sample to make inferences, which is also an important aspect to consider in probabilistic inference. The use of the strategy mentioned previously, the "total weight approach", was noticeable in the first two tasks (Task5Inst and Task6Inst), changing during the last task (Task7Inst). It seems that the format of the worksheets was driving their decision regarding sample size. The worksheets students used during the instruction included several tables they needed to fill out (see Appendix A). These tables asked them to run pre-determined number of trials, starting with 10 trials, 20 trials, and 30 trials. At the end there was a table where students had to choose the number of trials to run. Since they had already run samples of size 10, 20, and 30 most of the students decided to use a number larger than 30, such as 50.

During the last task students seems to have a clearer perception about the importance of sample size to make inferences. Only by analyzing the poster they created, it seems that their final decisions were more influenced by samples of large size. This observation was confirmed by Lee et al. (2009) as they stated that "instances where students employed a large sample size had substantial impact on [students] reasoning" (p.10). After instruction, the level of reasoning of three students stayed pretty consistent, but the level of reasoning for the other three students showed inconsistency between the different modes (interview and written tests).

Limitations

The first limitation of this study is that it is based on case study research making it difficult to make any generalization to bigger populations. However, the purpose of case study research is not to make larger generalizations, but to study in-depth how instruction could possibly impact students' informal probabilistic inference abilities. Even though, the results presented in this investigation may be true for other sixth grade students nationwide. Nonetheless, the framework developed during the study was based on key finding in literature. Therefore, it could be used to characterize students' ability to make probabilistic inferences in any situation.

Second, since most of the work analyzed was written, there were lots of blank spaces and, in some occasions, arguments were short and vague. This made it difficult to characterize students' level of reasoning because of the lack of responses. Variability was the most difficult construct to characterize using the framework given the lack of responses and vague arguments students offered at times.

Third, during the instructional tasks, there are videos that captured the image of the computer as they worked as well as videos with their conversations and negotiation of ideas. Given that the author did not use those videos, her analysis was based on students' answers to the worksheet they completed in class and the poster they presented to the class during the Schoolopoly task. The videos would help to get a better sense of what the reasoning behind their responses was.

Lee et al. (2009) reported results after an in depth analysis of the videos while the six case students were engaged working on the Schoolopoly task (Task7Inst). Their results showed that, although most of the cumulative sample sizes were100 or below, students' reasoning was most impacted by the samples of large size. Lara and Dannie reported in their poster that the school should buy dice from Dice R' Us because they were fair, but at the same time the estimation of the theoretical probabilities did not seem to belong to a fair die. In their article Lee et al. (2009) explained Lara was not convinced the dice were fair, but the social interaction with Dannie convinced her about the fairness of the dice. Also it seem that they interpreted Question 3 (see Figure 2.2) as to provide relative frequencies of empirical data, so Lara ran a new experiment of 36 trials and reported the empirical results in their poster. The results from the Schoolopoly task presented in Lee et al. (2009) about Manuel and Brandon seems to be consistent with the results found through this investigation. Their representation in the poster and the large data sets they used seems to direct them to reject their initial hypothesis that the dice were fair. The estimated theoretical probabilities were based on a sample of size 6000. Regarding Greg and Jasyn, their recognition about variability was noticeable in their poster and it is comparable with the results obtained by Lee et al. (2009).

Implications

Literature indicates that lack of recognition about the importance of sampling and variability are two of the biggest problems in statistics (Pfannkuch, 2005). With the design of

appropriate tasks, such as the ones described in Zieffler et al. (2009), which primarily seek to encourage thinking about variability, the probabilistic framework developed in this investigation could be used to characterize students' levels of reasoning regarding the recognition of variability.

Teachers should be informed with literature about how students learn probabilistic inference. They should attend professional development workshops where specialized people trained them on how to create meaningful tasks for their students. It is hard for students to understand important concepts such as variability (Garfield & Ben-Zvi, 2005, 2008; Hammerman & Rubin, 2004), and sampling (Pfannkuch, 2005) and as mentioned by Garfield and Ben-Zvi (2008) it is challenging to prepare teachers to create learning environments to develop a deep and meaningful understanding of statistics in their students. The results obtained through this investigation showed there are types of task that engage students more than others. Tasks such as the Mystery Bag of Marbles, Mystery Fish in a Lake, and the Schoolopoly tasks should be used to engage teachers in informal probabilistic inference so they experience they high reasoning necessary to complete those tasks and be eager to learn how to design those types of tasks for their students. As discussed by Cobb and McClain (2004), there are five principles all teachers should take into account when designing tasks driven to help students develop statistical reasoning. Those principles are essentially a guide for teachers to follow when designing tasks to enhance students reasoning of statistical inference. The five principles are:

- *First*, the task should be focused on developing only one central statistical idea instead of presenting a set of tools and procedures.
- *Second*, the teacher should use real and motivating data sets to engage students in making conjectures that are meaningful to them.
- *Third*, the teacher should determine the structure of the classroom activities more appropriate to help students develop their statistical reasoning (i.e. group vs. individual activities, whole class discussion).

- *Fourth*, the teacher needs to determine the appropriate technology tool that will allow students to explore the data and test their conjectures, and
- *Fifth* the teacher should promote classroom discourse focused on statistical arguments and significant statistical ideas.

As recommended by the *National Council of Teachers of Mathematics* (NCTM, 2000) "through the grades, students should be able to move from situations for which the probability of an event can readily be determine to situations in which sampling and simulations help them quantify the likelihood of an event" (p. 51). These types of situations were used during the instructional part of the study, where students used a simulation tool to collect samples and use the information collected to estimate the likelihood of an event. Tasks similar to the ones used to engage students in informal probabilistic inference during this study could be used in professional development workshops as well as in future research.

The GAISE framework describes statistical problem solving as an investigation process where students should formulate questions, collect data, analyze data, and interpret results. As teachers develop their ability to design meaningful tasks for their students, they should take into account the statistical problem solving framework proposed by GAISE. Tasks such as the ones used during this study presented situations where the question was already formulated, but students needed to collect the data, analyze it, and interpret the results in order to draw inferences supported by the data collected.

Although the Probabilistic Inference framework described during this study was not easy to develop, the validation process gave the researcher insights about how middle school students reason about informal probabilistic inference, what type of tasks are meaningful for them, and what type of classroom arrangements are most appropriate for them to work and reason at higher levels. The Probabilistic Inference framework could be use as a tool to study how students engage in informal probabilistic inference in many settings.

Recommendation for Future Research and Conclusion

For future research the author recommends the use of the Probabilistic Inference framework she developed to characterize students' reasoning about probabilistic inference. This will help the author to refine the framework even more. Also for future research the author is interested in analyzing the videos taken during instruction that include the social interaction of the pairs as well as an image of their computer screen while they were engaged in the Mystery Bag of Marbles task and the Mystery Fish on a Lake task. The videos related

to the Schoolopoly task have been analyzed and results have been reported in Lee et al. (2009).

This research included a total of 23 students, but only six were selected as case-study. All 23 students took the pre-test, post-test, and the retention test which make possible to conduct a statistical analysis to investigate whether or not there were significant changes in levels of reasoning for all students from the pre-test to the post-test and from the post-test to the retention test, as well as from the pre-test to the retention test.

The results indicated that, in general, students' ability to engage in probabilistic inferences grew over the course of instruction, although their levels of reasoning were inconsistently applied after instruction. If the research will be conducted again in order to collect more data to validate the framework even more, it would be good to have students engage with the instructional tasks for a prolonged period of time, working in groups and with access to a simulation tool, and observe if the time exposed to instruction has an effect on the consistency of students' levels of reasoning after instruction.

The tasks used over the course of the instruction during this study could be used at the elementary level to measure student's level of reasoning on the three constructs (generalization, variability, and investigation) and see whether elementary school students' reasoning is affected by the mode of engagement of the tasks, as result indicated during this study with middle school students.

The research in students' reasoning about probabilistic inference is nascent, and more

research is needed in the teaching and learning of probabilistic inference. Shaughnessy (1992, 2007) and Jones et al. (2007) have called for more research on how students use data to make inferences about an underlying population. They also have called for the necessity of teaching experiments in probability in order to determine the impact instruction has on students' informal statistical inference. Results from this study have helped to answer both calls and aimed to collaborate with the mathematical education community in understanding how students reason about data and how they make inference about and underlying population.

Results from this study suggest that students' ability to make probabilistic inferences increased over the course of instruction. It suggested that students' level of reasoning about generalization, variability, and investigation are higher when students are engaged in tasks where they work with a partner and have access to a simulation tool. It seems the social interaction and all social negotiation and communication within partners are essential in the developing of higher levels of reasoning. It is important to acknowledge what type of instruction is beneficial for students to develop probabilistic reasoning and what type of tasks are more appropriate to reveal students' probabilistic thinking in order for researchers to be able to construct appropriate curricular materials to foster students' informal probabilistic inference and to be informed to design effective professional development for teachers (Zieffler et al., 2008).

The Probabilistic Inference framework developed during this study can be used in

future research on statistics education in order to further understand how students' reason about important aspects on probabilistic inference, such as, generalization beyond data, variability, and the investigation process.

Assessing students' informal probabilistic inference levels of reasoning is complicated and strongly depends on the mode of engagement the students are using to engage in the tasks. Different modes of engagement (e.g., interview, written test, and work with peers) offer students different ways of demonstrating understanding to their teacher. Teachers should take advantage of that and use the most appropriate mode of engagement in order to help students develop the abilities they need to make statistical inference. Teachers could use the Probabilistic Inference framework developed through this study to gather information about students' levels of reasoning in the three areas mentioned (generalization, variability, and investigation). Tasks such as the ones used through the 12-day instructional program should be used by prospective and practicing teachers to engage students in meaningful activities where students are asked to make inferences about context that interests them.

BIBLIOGRAPHY

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

Baxter, P., & Jack, S. (2008). Qualitative Case Study Methodology: Study Design and Implementation for Novice Researchers. *The qualitative report*, 13(4), 544-559.

Ben-Zvi, D. (2004). Reasoning about variability in comparing distributions. *Statistics Education Research Journal*, 3(2), 42-63

Ben-Zvi, D. (2006). Scaffolding students' informal inference and argumentation. In A. Rossman and B. Chance (Eds.), *Working Cooperatively in Statistics Education. Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. Voorburg, The Netherlands: International Statistical Institute. [Online: http://www.stat.auckland.ac.nz/~iase/publications/17/2D1_BENZ.pdf]

Ben-Zvi, D., & Amir (2003). How do primary school students begin to reason about distributions? In K. Makar (Ed.), *Reasoning about Distributions: A collection of studies. Proceeding of the Fourth International Research Forum on Statistical Reasoning, Thinking and Literacy* (SRTL-4). Brisbane, Australia: University of Queensland.

Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi and J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (p. 3-16). Dordretch, The Netherlands: Kluwer Academic Publishers.

Berenson, S. (1999). Students' representation and trajectories of probabilistic thinking. In R. Hitt and M. Santos (Eds.), *Proceedings of the twenty first annual meeting of the North American chapter of the international group for the psychology of education*, Columbus, OH: ERIC Clearinghouse of Science, Mathematics, and Environmental Education.

Biggs, J.B., & Collis, K.F. (1982). *Evaluating the quality of learning: The SOLO Taxonomy*. New York, NY: Academic Press.

Biggs, J.B., & Collis, K.F. (1991). Multimodal learning and the quality of intelligent behavior. In H. Rowe (Ed.), *Intelligence, reconceptualization and measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Blee, K.M., & Taylor, V (2002). Semi-structured interviewing in social movement research. In B. Klandermans and S. Staggenborg (Eds.), *Methods of social movement research*, Minneapolis, MN: University of Minnesota Press.

Canada, D. (2006). Elementary pre-service teachers' conceptions of variability in a probability context. *Statistics Education Research Journal*, 5(1), 36-63.

[Available online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ_5(1)_Canada.pdf]

Chance, B., Ben-Zvi, D., Garfield, J., & Medina, E. (2007). The role of technology in improving student learning if statistics. Technology Innovations in Statistics Education Journal, 1(1), 1-24.

Cobb, P., & McClain, K. (2004). Principles of instructional design for supporting the development of students' statistical reasoning. In D. Ben-Zvi and J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*. The Netherlands: Kluwer Academic.

Drier, H.S. (2000). Children's meaning-making activity with dynamic multiple representations in a probability microworld. In M. Fernandez (Ed.), *Proceedings of the 22nd Annual Meeting of the North American chapter of the International Group for the Psychology of Mathematics Education*. Columbus, OH: ERIC Clearinghouse of Science, Mathematics, and Environmental Education.

Finzer, W. (2007). *Fathom Dynamic Data Software: Exploring Statistics with Fathom*. Emeryville, CA: Key Curriculum Press.

Fischbein, E., & Schnarch, E. (1997). The evolution with age of probabilistic, intuitively based misconceptions. *Journal for Research in Mathematics Education*, 28(1), 96-105.

Fischbein, E., Nello, M.S., & Marino, M.S. (1991). Factors affecting probabilistic judgments in children in adolencence. *Educational Studies in Mathematics*, 22, 523-549.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2005). Guidelines for assessment and instruction in statistics education (GAISE) report: A Pre-K-12 curriculum framework. Alexandria, VA: American Statistical Association. [Available online: http://www.amstat.org/education/gaise/]

Garfield, J., & Ben-Zvi, D. (2005). A framework for teaching and assessing reasoning about variability. *Statistics Education Research Journal*, 4(1), 92-99.

Garfield, J., & Ben-Zvi, D. (2008). Preparing school teachers to develop students' statistical reasoning. In C. Batanero, G. Burrill, C. Reading and A. Rossman (Eds.), *Joint ICMI/IASE Study: Teaching statistics in school mathematics. Challenges for teaching and teacher education. Proceedings of the ICMI study 18 and 2008 IASE round table conference.*

Hammerman, J., & Rubin, A. (2004). Strategies for managing statistical complexity with new software tools. Statistics Education Research Journal, 3(2), 17-41 [Available online: http://www.stat.auckland.ac.nz/serj]

Huff, D (1954). *How to lie with statistics*. Norton and Company, Inc: New York

Jones, G.A., Langrall C.W., & Mooney E.S. (2007). Research in probability: Responding to classroom realities. In F.K. Lester (Ed.), *The Second Handbook of Research on*

Mathematics (pp. 909-956). Reston, VA: National Council of Teachers of Mathematics.

Jones, G. A., Langrall, C. W., Thornton, C. A., & Mogill, A. T. (1997). A framework for assessing and nurturing young children's thinking in probability. *Educational Studies in Mathematics*, 32, 101-125.

Jones, G.A., Langrall C.W., Thornton, C.A. & Mogill, A.T. (1999). Student probabilistic thinking in instruction. *Journal for Research in Mathematics Education*, 30, 487-519

Konold, C., & Kazak, S. (2008) "Reconnecting Data and Chance", *Technology Innovations in Statistics Education*, 2(1).
[Available online: http://repositories.cdlib.org/uclastat/cts/tise/vol2/iss1/art1]

Konold, C. & Miller, C. (2004). *TinkerPlots: Dynamic Data Explorations*. Emeryville, CA: Key Curriculum Press.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signal in noisy processes. *Journal for Research in Mathematics Education*, 33(4), 259-289.

Lee, H.S. (2005). Students' reasoning with small and large trials in probability simulations. In Lloyd, G. M., Wilson, M., Wilkins, J. L. M., & Behm, S. L. (Eds.), *Proceedings of the 27th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* [CD Room]. Eugene, OR: All Academic.

Lee H. S., Angotti, R., & Tarr, J. (2010) Making comparisons between observed data and expected outcomes: Students' informal hypothesis testing with probability simulation tools. *Statistics Education Research Journal 9*(1), 68-96. [Available online: http://www.stat.auckland.ac.nz/serj].

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, 8(1), 82-105. [Available online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ8(1)_Makar_Rubin.pdf]

National Council of Teachers of Mathematics (1989). *Principles and Standards: Curriculum and Evaluation*. Reston, VA: The Council.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: The Council.

Paparistodemou, E., & Meletiou-Mavrotheris, M. (2008). Developing young students' informal inference skills in data analysis. *Statistics Education Research Journal*, 7(2), 83-106.

Pegg, J., & Devey, G. (1998). Interpreting student understanding in Geometry: A synthesis of two Models. In R. Lehrer and D. Chazan (Eds.), *Designing learning environments for developing understanding of geometry and space* (pp. 109-135). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Pfannkuch, M. (2006a). Comparing box plot distributions: a teacher's reasoning. *Statistics Education Research Journal*, 5(2), 27-45. [Available online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ5(2)_Pfannkuch.pdf]

Pfannkuch, M. (2006b). Informal inference reasoning. In A. Rossman and B. Chance (Eds.), *Working Cooperatively in Statistics Education. Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. Voorburg: The Netherlands: International Statistical Institute. [Available online: http://www.stat.auckland.ac.nz/~iase/publications/17/6A2_PFAN.pdf]

Pratt, D. (2000). Making sense of the total of two dice. *Journal of Research in Mathematics Education*, 31, 602-625.

Pratt, D. (2005). How do teachers foster students' understanding of probability? In Graham A. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning*

Pratt, D., Johnston-Wilder, P., Ainley, J., & Mason, J. (2008). Local and global thinking on statistical inference. *Statistics Educational Research Journal*, 7(2), 107-129. [Available online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ7(2)_Pratt.pdf]

Rubel, L. H. (2006). Students' probabilistic thinking revealed: The case of coin tosses. In G. Burrill (Ed.), *Thinking and Reasoning with Data and Chance: Sixty-eighth Yearbook* (pp. 49-59). Reston, VA: National Council of Teachers of Mathematics.

Rubin, A., Hammerman, J., & Konold, C. (2006). Exploring informal inference with interactive visualization software. In A. Rossman and B. Chance (Eds.), *Working Cooperatively in Statistics Education. Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. Voorburg: The Netherlands: International Statistical Institute. [Available online: http://www.stat.auckland.ac.nz/~iase/publications/17/2D3_RUBI.pdf]

Sedlmeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making*, 10, 33-51.

Shaughnessy, M. (1992). Research on probability and statistics: Reflections and directions. In D. Grouws (Ed.). *Handbook of Research on Mathematics Teaching and learning* (pp. 465-494). Reston, VA: National Council of Teachers of Mathematics.

Shaughnessy, M. (2007). Research on statistical learning and reasoning. In F.K. Lester (Ed.), *The Second Handbook of Research on Mathematics* (pp. 957-1008). Reston, VA: National Council of Teachers of Mathematics.

Sorto, M.A. (2006). Indentifying content knowledge for teaching statistics. In A. Rossman and B. Chance (Eds.), *Working Cooperatively in Statistics Education. Proceedings of*

the Seventh International Conference on Teaching Statistics, Salvador, Brazil. Voorburg: The Netherlands: International Statistical Institute. [Available online: http://www.stat.auckland.ac.nz/~iase/publications/17/C130.pdf]

Stake, R. E. (1995). *The art of case study research*. Thousand Oaks, CA: Sage.

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

Stohl, H., & Tarr, J.E. (2002b). Using multi-representational computer tools to make sense of inference. In, D. Mewborn (Ed.), *Proceedings of the twenty-fourth annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Athens, GA: Columbus, OH: ERIC Clearinghouse for Science Mathematics and Environmental Education.

Stohl, H. (1999-2005). Probability Explorer. Software application distributed by author at http://www.probexplorer.com.

Tarr, J. E., Lee, H. S., & Rider, R. L. (2006). When data and chance collide: Drawing inferences from empirical data. In G. Burrill (Ed.), *Thinking and Reasoning with Data and Chance: Sixty-eighth Yearbook* (pp. 139-150). Reston, VA: National Council of Teachers of Mathematics.

Tukey, J. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Vallecillos, A., & Moreno, A. (2002). Framework for instruction and assessment on elementary inferential statistics thinking. Presentation at the Second International Conference on the Teaching of Mathematics, Crete, Greece, July 1-6.

Watson, J. (2001). Longitudinal development of inferential reasoning by school students. *Educational Studies in Mathematics*, 47, 337-372.

Watson, J. (2008). Exploring beginning inference with novice grade 7 students. *Statistics Education Research Journal*, 7(2), 59-82. [Available online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ7(2)_Watson.pdf]

Watson, J., Collis, K., & Mortiz, J. (1997). The development of chance measurement. *Mathematics Education Research Journal, 9*, 60-82.

Watson, J., & Donne, J. (2008). Building informal inference in grade 7. In M. Goos, R. Brown, & K. Makar (Eds.), *Proceedings of the 31st Annual Conference of the Mathematics Education Research Group of Australasia*, St Lucia, Brisbane, Queensland, Australia. (In Press)

Watson, J., & Donne, J. (2009). *TinkerPlots* as a researcher tool to explore students understanding. *Technology Innovation in Statistics Education*, 3(1), 1-34. [Available online: http://repositories.cdlib.org/uclastat/cts/tise/vol3/iss1/art1/]

Yin, R.K. (2003). *Case study research: design and methods* (3rd ed.). Thousand Oaks, CA: Sage.

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

APPENDICES

Task1PreT
(Buffalo Nickel)

Kiki found an old buffalo nickel on the side of the street. She wondered if this coin would be fair to use for the daily coin toss with her brother to decide who takes out the trash. Kiki tossed the coin 10 times and got 7 tails and 3 heads. Based on this data, is this a fair coin to use? Why or why not?

Kiki continued tossing the coin until she had 100 tosses and got 67 heads and 33 tails. Based on this data, is this a fair coin to use? Why or why not?

Again, Kiki continued tossing the coin until she had 1000 tosses and got 643 tails and 357 heads. Based on this data, is this a fair coin to use? Why or why not?

Task2PreT
(Buffalo Nickel, cont.)

If you found an old coin and needed to estimate the probability of the coin landing on heads, what would you do? What would give you convincing evidence?

Task3PreI
(Mystery Bags of Marbles)

SAY: Shown here is a bag containing 10 marbles. Only 2 colors are present in Bag A: blue marbles and red marbles.

DO: (Show student Card 4)

SAY: Suppose I were to reach into the bag without looking and select a single marble, look at it, record its color, and return it to the bag. How many times would I need to draw out a single marble – replacing it after each draw – in order to be *very confident* about how many marbles of each color are in Bag A? Explain.

SAY: (If they indicate 10 trials is sufficient) Suppose we got 1 blue and 9 red marbles. Would that tell you about the contents of the bag?

SAY: Shown here is a bag containing 10 marbles. There are 4 colors are present in Bag B: blue marbles, red marbles, green marbles and yellow marbles.

DO: (Show student Card 4)

SAY: Suppose I were to reach into the bag without looking and select a single marble, look at it, record its color, and return it to the bag. How many times would I need to draw out a single marble – replacing it after each draw – in order to be *very confident* about how many marbles of each color are in Bag B? Explain.

Task4PreI
(The Plastic Cup)

DO: Have the student examine a plastic cup.

SAY: While setting out plastic cups for a potluck (picnic), you drop several of them on the floor. You notice that some of the cups landed on their side, others landed right side up and still others landed upside down. Suppose you wanted to estimate the likelihood that the cup would land upside down. How many times would you need to drop a cup in order to be confident in your estimate?

Task5Inst
(Mystery Bag of Marbles)

There is a bag of marbles hidden that contains 10 marbles. We will do some experimentation to predict the contents of the bag. We will choose 1 marble at a time with replacement from the bag.

Use Probability Explorer to run **10** trials (picking a marble with replacement). Repeat this 6 times, clearing your data after each set of 10. Record data in the table below

Number of trials	White	Green	Red	Black	Yellow	Blue
1st 10						
2nd 10						
3rd 10						
4th 10						
5th 10						
6th 10						

1. Based on data you collected, predict how many of each color you think are in the bag.
2. How confident are you that your prediction is accurate? Why?

Use Probability Explorer to run **20** trials (picking a marble with replacement). Be sure to save an image of either the bar or pie graph in your notebook. Repeat this 6 times, clearing your data after each set of 20. Record the data in the table below.

Number of trials	White	Green	Red	Black	Yellow	Blue
1st 20						
2nd 20						
3rd 20						
4th 20						
5th 20						
6th 20						

1. Based on data you collected, predict how many of each color you think are in the bag.
2. How confident are you that your prediction is accurate? Why?

Clear all your data. With your partner, use the tools in Probability Explorer to run some more simulations and collect data that would help you make your ***best prediction*** of the contents in the bag. Record your data below.

Number of trials	White	Green	Red	Black	Yellow	Blue

1. Based on the data you collected, what is your BEST prediction for how many of each color is in the bag.
2. Why is your BEST prediction?

In pairs, you will get to design your own "guess my bag" simulation. Open the bag of marbles. Clear all marbles. Have one partner close their eyes while the other partner fills the bag with marbles.

The bag must have **12** marbles in it, but can have either 2 or 3 colors of your choice. After partner A designs the mystery bag, partner B must use the tools in Probability Explorer to try to predict what is in the bag of marbles. Record all data in the chart below.

Who designed the bag? _____ (partner A)
Who is trying to "Guess My Bag"? _____ (partner B)

Number of trials	White	Green	Red	Black	Yellow	Blue

1. Based on data collected what is the BEST prediction for what is in the bag?
2. Why is this your best prediction? Explain your reasoning for your prediction.
3. What were the contents in the bag designed by the first partner?

Task6Inst
(Mystery Fish in Lake)

There is a lake nearby that has two types of fish in it, Blue Bass and Green Gills. The lake was just stocked with fish but no one knows how many fish are in the lake or the number of Blue Bass and Green Gills. There is a fishing contest coming up soon. The winner will be declared based on who catches the most Blue Bass. Assuming that a contestant catches a fish, the sponsors of the contest need to know the probability of that fish being a Blue Bass.

The Probability Explorer has been set up to model the lake containing Blue Bass and Green Gills. You need to simulate catching fish, record and analyze the data, and estimate the probability that a fish is a Blue Bass. Since you can NOT keep fish out of water, every time you catch a fish, you have to throw it back into the lake. Thus, you are doing an experiment *with replacement*.

Use Probability Explorer to run **10** trials. Save a copy of either a pie or bar graph in the Notebook. Repeat this 6 times, clearing the data after each set of 10. Record data in the table

Number of trials	Blue Bass	Green Gills
1st 10		
2nd 10		
3rd 10		
4th 10		
5th 10		
6th 10		

1. Based on the data you collected, what would be your estimate of the probability that a fish caught is a Blue Bass? Explain your reasoning

Use Probability Explorer to run **30** trials. Be sure to save an image of either the bar or pie graph in your notebook. Repeat this 6 times, clearing your data after each set of 30. Record the data in the table below.

Number of trials	Blue Bass	Green Gills
1st 30		
2nd 30		
3rd 30		
4th 30		
5th 30		
6th 30		

1. Based on the data you collected for the sets of 30 trials, estimate the probability that a fish caught is a Blue Bass.
2. How did you use your data to make your estimate?
3. Did your estimate change from Part I when you ran sets of 10 trials? Why or why not?

Clear all your data. With your partner, use the tools in Probability Explorer to run some more simulations and collect data that would help you make your best estimate of the probability that a fish caught is a Blue Brass. Record the data below.

Number of trials	Blue Bass	Green Gills

Task7Inst
(Schoolopoly)

Your school is planning to create a board game modeled on the classic game of *Monopoly*^{TM}. The game is to be called *Schoolopoly* and, like *Monopoly*^{TM}, will be played with dice. Because many copies of the game expect to be sold, companies are competing for the contract to supply dice for *Schoolopoly*. Some companies have been accused of making poor quality dice and these are to be avoided since players must believe the dice they are using are actually "fair." Each company has provided a sample die for analysis and you will be assigned one company to investigate:

Luckytown Dice Company	Dice, Dice, Baby!
Dice R' Us	Pips and Dots
High Rollers, Inc.	Slice n' Dice

Your Assignment

Working with your partner, investigate whether the die sent to you by the company is, in fact, fair. That is, are all six outcomes equally likely to occur? You will need to create a poster to present to the School Board. The following three questions should be answered on your poster:

1. Would you recommend that dice be purchased from the company you investigated?

2. What evidence do you have that the die you tested is fair or unfair?

3. Use your experimental results to estimate the theoretical probability of each outcome, 1-6, of the die you tested.

Use Probability Explorer to collect data from simulated rolls of the die. Copy any graphs and screen shots you want to use as evidence and paste them in a Word document. Later, you will be able to print these.

Task8PostT
(Bottle Cap)

Alicia found a bottle cap on the sidewalk. She told her brother Marcus, "Instead of using a coin, we could flip this bottle cap to decide who gets to use the computer first when they get home." They took turns flipping the bottle cap.

After 10 tosses, the bottle cap landed up 7 times and down 3 times. Based on this data, is the bottle cap fair to use? Why or why not?

After 100 tosses, the bottle cap landed up 63 times and down 37 times. Based on this data, is the bottle cap fair to use? Why or why not?

After 1000 tosses, the bottle cap landed up 649 times and down 351 times. Based on this data, is the bottle cap fair to use? Why or why not?

Task9PostT
(Bottle Cap, cont.)

If you found a bottle cap and needed to determine if it was fair or not, what would you do? What would give you convincing evidence?

Task10PostI
(Mystery Jars of Jolly Ranchers)

SAY: Shown here is a jar containing 10 Jolly Ranchers. Only 2 flavors are present in Jar A: Cherry and Grape.

DO: (Show student Card 4)

SAY: Suppose I were to reach into the jar without looking and select a single Jolly Rancher candy, look at it, record its flavor, and return it to the jar. How many times would I need to draw out a single piece of candy – replacing it after each draw – in order to be *very confident* about how many candies of each flavor are in Jar A? Explain.

SAY: (If they indicate 10 trials is sufficient) Suppose that we got 1 cherry and 9 grape Jolly Rancher candies. Would that tell you about the contents of the jar?

SAY: Shown here is a jar containing 10 Jolly Rancher candies. There are four flavors present in Jar B: cherry, grape, apple, and lemon.

DO: (Show student Card 4)

SAY: Suppose I were to reach into the jar without looking and select a single Jolly Rancher candy, look at it, record its flavor, and return it to the jar. How many times would I need to draw out a single candy – replacing it after each draw – in order to be *very confident* about how many candies of each flavor are in Jar B? Explain.

Task11PostI
(The Toothpaste Cap)

DO: Have the student examine a toothpaste cap.

SAY: While brushing your teeth one morning, you drop the toothpaste cap on the floor. You notice that it landed on its side and figured that it also could have landed right side up or upside down. Suppose you wanted to estimate the likelihood that the toothpaste cap would land upside down. How many times would you need to drop the toothpaste cap in order to be confident in your estimate?

Task12RetT
(Shaped button)

Nathalie found an unusually shaped button while playing at the park. On side of the button is flat and the other side has several ridges on it. She told her brother Steven, "If I toss this button, I think it's more likely to land with the ridges side up!" They took turns flipping the button.

After 10 tosses, the button landed with its ridged side up 7 times and down 3 times. Based on this data, do you think the button is actually more likely to land with the ridged side up? Why or why not?

After 100 tosses, the button landed ridged side up 65 times and down 35 times. Based on this data, do you think the button is actually more likely to land with the ridged side up?? Why or why not?

After 1000 tosses, the button landed ridged side up 627 times and down 373 times. Based on this data, do you think button is actually more likely to land with the ridged side up? Why or why not?

Task13RetT
(Shaped Button, cont.)

Suppose you found an unusually shaped button that was flat on one side and had ridges on the other side. You were curious if the button was more likely to land on either of the sides when flipped. What would you do to determine if it was more likely to land on a particular side? What would give you convincing evidence that the button was more likely to land one way?

Table B.1. Probabilistic Inference Framework

	Pre-structural	**Uni-structural**	**Multi-structural**	**Relational**
Generalization (G)	• Makes generalizations based on their previous experiences and their intuitions. Ignores sample size.	• Makes generalizations based on a single part of the information given ignoring the others.	• Makes generalizations based on most of the given information or data collected but still ignoring others.	• Make generalizations based on all the information given or the data collected. Recognizes the importance of large samples to make generalizations.
Variability (V)	• Attributes variability to luck or chance in all contexts. No recognition that variability can be described/controlled.	• Recognizes variability among individual results within a single sample (e.g., I got more 5's than 3's). • Has low expectation of variability from an expected distribution based on some theoretical probability distribution.	• Recognizes variability among individual results within a single sample and recognizes variability across different samples (e.g., last time I got more 5's, this time I got more 1's). • Has a better sense of expecting some variability from theoretical probability distributions, but may expect too much variability.	• Recognizes variability within and across samples and relates variability to characteristics of the underlying probability distribution.
Investigation (I)	• Does not propose a data collection process • Believes sampling does not tell you anything about the underlying population.	• Proposes a data collection process, either appropriate or inappropriate (e.g., uses a small sample size), and does not describe how to assess the results.	• Proposes an appropriate or inappropriate data collection process and describes how the results could be analyzed, although not using the most appropriate approach.	• Proposes an appropriate data collection process (e.g., a large sample size) and describes how the results will be analyzed (e.g., from the distribution of the outcomes)

Table B.2. Probabilistic Inference Framework used for Sequence 1

	Pre-structural	**Uni-structural**	**Multi-structural**	**Relational**
Generalization (G)	• Generalize the button, the cap, or the nickel is fair or unfair based in their intuitions or physical appearances.	• Generalize the button, the cap, or the nickel is fair or unfair using a single part of the information given. • Make generalizations using a small sample, does not recognize the importance of sample size to make generalizations.	• Generalize the button, the cap, or the nickel is unfair and mentioned that the numbers should be closer together. • Use more that a single part of information, but does not connect them.	• Generalize the button, the cap, or the nickel is unfair and mentioned that the numbers should be closer together and recognize the importance of sample size (e.g., 10 tosses is not enough to determine if it is fair or not).
Variability (V)	• Attributes variability to luck or chance in all contexts. No recognition that variability can be described/controlled.	• Recognizes variability among individual results within a single sample (e.g., there are more heads than tails).	• Recognizes variability among individual results within a single sample and recognizes variability across different samples (e.g., recognizes there are more heads in each sample).	• Recognizes variability within and across samples and relates variability to characteristics of the underlying probability distribution.
Investigation (I)	• Does not apply	• Does not apply	• Does not apply	• Does not apply

Table B.3. Probabilistic Inference Framework used for Sequence 2

	Pre-structural	**Uni-structural**	**Multi-structural**	**Relational**
Generalization (G)	• Make generalizations about the fairness of the button, the cap, or the nickel based in their intuitions or physical appearances.	• Makes generalizations about the fairness of the button, the cap, or the nickel based on the results from a single small or large sample.	• Makes generalizations about the fairness of the button, the cap, or the nickel based on results from several sample, wither small or large. • Does not make connections about sample size and generalization.	• Makes generalizations about the fairness of the button, the cap, or the nickel based on results from several samples of appropriate size (e.g., samples of size 100 or more, or a single sample of size 1000 or more).
Variability (V)	• Does not recognize variability can be described/controlled.	• Recognizes variability among individual results within a single sample. • Has low expectation of variability from an expected distribution based on some theoretical probability distribution	• Recognizes variability among individual results within a single sample and recognizes variability across different samples. • Has a better sense of expecting some variability from theoretical probability distributions, but may expect too much variability.	• Recognizes variability within and across samples and relates variability to characteristics of the underlying probability distribution.
Investigation (I)	• Does not propose a data collection process.	• Proposes an inappropriate data collection process, and does not describe how to assess the results (e.g., propose a small sample size)	• Propose an appropriate or inappropriate data collection process, and • Describe an inappropriate way to assess the results.	• Propose an appropriate data collection process and describe how to assess the results (e.g., uses the distribution of the outcomes to analyze the result).

Table B.4. Probabilistic Inference Framework used for Sequence 3

	Pre-structural	**Uni-structural**	**Multi-structural**	**Relational**
Generalization (G)	• Does not make generalizations about the content of the bag or the jar.	• Make generalizations based on samples of small sizes and does not use the distribution of the outcomes to make generalizations about the content of the bag or the jar. • Uses the result from a small sample as the prediction of the content of the bag or jar (e.g., from 10 trials if you got 1 blue and 9 red, then there is 1 blue and 9 reds).	• Make generalizations based on samples of small/large sizes and uses the distribution of the outcomes to make generalizations about the content of the bag or the jar. • Does not use results from small samples to make generalization about the content of the bag or the jar (e.g., from 10 trials if you got 1 blue and 9 red, then you probably have more reds than blues). • Although consider sample size, does not connect it with its importance to make generalizations.	• Make generalization on samples of large size and does use the distribution of the outcomes to make generalizations. • Recognizes the importance of large samples to make generalizations.
Variability (V)	• Attributes variability to luck or chance in all contexts. • No recognition that variability can be described/controlled.	• Recognizes variability among individual results within a single sample (e.g., there are more reds than blues) or recognize variability across different samples, but not both.	• Recognizes variability among individual results within a single sample and recognizes variability across different samples (e.g., last time I got more reds, now I got more blues).	• Recognizes variability within and across samples and relates variability to characteristics of the underlying probability distribution.
Investigation (I)	• Does not propose a data collection process.	• Propose a data collection process, either appropriate or inappropriate, and does not propose how to assess the results (e.g., proposes a small sample size)	• Propose a data collection process, either appropriate or inappropriate, and does not reasons appropriately on how to assess the results.	• Propose an appropriate data collection and reasons appropriately on how to assess the results.

Table B.5. Probabilistic Inference Framework used for Sequence 4

	Pre-structural	**Uni-structural**	**Multi-structural**	**Relational**
Generalization (G)	• Does not make generalizations about the likelihood of the plastic cup, the fish at the lake, the dice, or the toothpaste cap.	• Make generalizations based on samples of small sizes and does not use the distribution of the outcomes to make generalizations. • Uses the result from a small sample as representative of the underlying population (e.g., if a toothpaste cap landed upside down 2 times out of 10, then the probability that the toothpaste cap landed upside down is 20%).	• Make generalizations based on samples of small/large sizes and uses the distribution of the outcomes to make generalizations. • Could consider sample size, but does not connect it with its importance to make generalizations.	• Make generalization on samples of large size and does use the distribution of the outcomes to make generalizations. • Recognizes the importance of large samples to make generalizations.
Variability (V)	• Attributes variability to luck or chance in all contexts. • No recognition that variability can be described/controlled.	• Recognizes variability among individual results within a single sample (e.g., there are more reds than blues) or recognize variability across different samples, but not both. • Has low expectation of variability from an expected distribution based on some theoretical probability distribution.	• Recognizes variability among individual results within a single sample and recognizes variability across different samples (e.g., last time I got more reds, now I got more blues). • Has a better sense of expecting some variability from theoretical probability distributions, but may expect too much variability	• Recognizes variability within and across samples and relates variability to characteristics of the underlying probability distribution.
Investigation (I)	• Does not propose a data collection process.	• Propose a data collection process, either appropriate or inappropriate, and does not propose how to assess the results (e.g., proposes a small sample size).	• Propose a data collection process, either appropriate or inappropriate, and does not reasons appropriately on how to assess the results.	• Propose an appropriate data collection and reasons appropriately on how to assess the results.

Table C.1. Tasks included in the 1st sequence

Task1PreT	Task8PostT	Task12RetT
Kiki found an old buffalo nickel on the side of the street. She wondered if this coin would be fair to use for the daily coin toss with her brother to decide who takes out the trash. Kiki tossed the coin 10 times and got 7 tails and 3 heads. Based on this data, is this a fair coin to use? Why or why not? Kiki continued tossing the coin until she had 100 tosses and got 67 heads and 33 tails. Based on this data, is this a fair coin to use? Why or why not? Again, Kiki continued tossing the coin until she had 1000 tosses and got 643 tails and 357 heads. Based on this data, is this a fair coin to use? Why or why not?	Alicia found a bottle cap on the sidewalk. She told her brother Marcus, "Instead of using a coin, we could flip this bottle cap to decide who gets to use the computer first when they get home." They took turns flipping the bottle cap. After 10 tosses, the bottle cap landed up 7 times and down 3 times. Based on this data, is the bottle cap fair to use? Why or why not? After 100 tosses, the bottle cap landed up 63 times and down 37 times. Based on this data, is the bottle cap fair to use? Why or why not? After 1000 tosses, the bottle cap landed up 649 times and down 351 times. Based on this data, is the bottle cap fair to use? Why or why not?	Nathalie found an unusually shaped button while playing at the park. On side of the button is flat and the other side has several ridges on it. She told her brother Steven, "If I toss this button, I think it's more likely to land with the ridges side up!" They took turns flipping the button. After 10 tosses, the button landed with its ridged side up 7 times and down 3 times. Based on this data, do you think the button is actually more likely to land with the ridged side up? Why or why not? After 100 tosses, the button landed ridged side up 65 times and down 35 times. Based on this data, do you think the button is actually more likely to land with the ridged side up?? Why or why not? After 1000 tosses, the button landed ridged side up 627 times and down 373 times. Based on this data, do you think button is actually more likely to land with the ridged side up? Why or why not?

Table C.2. Tasks included in the 2nd sequence

Task2PreT	Task9PostT	Task13RetT
If you found an old coin and needed to estimate the probability of the coin landing on heads, what would you do? What would give you convincing evidence?	If you found a bottle cap and needed to determine if it was fair or not, what would you do? What would give you convincing evidence?	Suppose you found an unusually shaped button that was flat on one side and had ridges on the other side. You were curious if the button was more likely to land on either of the sides when flipped. What would you do to determine if it was more likely to land on a particular side? What would give you convincing evidence that the button was more likely to land one way?

Table C.3. Tasks included in the 3rd sequence

Task3PreI	Task5Inst	Task10PostI
SAY: Shown here is a bag containing 10 marbles. Only 2 colors are present in Bag A: blue marbles and red marbles. DO: (Show student Card 4) SAY: Suppose I were to reach into the bag without looking and select a single marble, look at it, record its color, and return it to the bag. How many times would I need to draw out a single marble – replacing it after each draw – in order to be *very confident* about how many marbles of each color are in Bag A? Explain. SAY: (If they indicate 10 trials is sufficient) Suppose we got 1 blue and 9 red marbles. Would that tell you about the contents of the bag? SAY: Shown here is a bag containing 10 marbles. There are 4 colors are present in Bag B: blue marbles, red marbles, green marbles and yellow marbles. DO: (Show student Card 4) SAY: Suppose I were to reach into the bag without looking and select a single marble, look at it, record its color, and return it to the bag. How many times would I need to draw out a single marble – replacing it after each draw – in order to be *very confident* about how many marbles of each color are in Bag B? Explain.	There is a bag of marbles hidden that contains 10 marbles. We will do some experimentation to predict the contents of the bag. We will choose 1 marble at a time with replacement from the bag. Use Probability Explorer to run 10 trials (picking a marble with replacement). Repeat this 6 times, clearing your data after each set of 10. 1. Based on data you collected, predict how many of each color you think are in the bag. 2. How confident are you that your prediction is accurate? Why? Use Probability Explorer to run 20 trials (picking a marble with replacement). Be sure to save an image of either the bar or pie graph in your notebook. Repeat this 6 times, clearing your data after each set of 20. 1. Based on data you collected, predict how many of each color you think are in the bag. 2. How confident are you that your prediction is accurate? Why? Clear all your data. With your partner, use the tools in Probability Explorer to run some more simulations and collect data that would help you make your *best prediction* of the contents in the bag. 1. Based on the data you collected, what is your BEST prediction for how many of each color is in the bag. 2. Why is your BEST prediction?	SAY: Shown here is a jar containing 10 Jolly Ranchers. Only 2 flavors are present in Jar A: Cherry and Grape. DO: (Show student Card 4) SAY: Suppose I were to reach into the jar without looking and select a single Jolly Rancher candy, look at it, record its flavor, and return it to the jar. How many times would I need to draw out a single piece of candy – replacing it after each draw – in order to be *very confident* about how many candies of each flavor are in Jar A? Explain. SAY: (If they indicate 10 trials is sufficient) Suppose that we got 1 cherry and 9 grape Jolly Rancher candies. Would that tell you about the contents of the jar? SAY: Shown here is a jar containing 10 Jolly Rancher candies. There are four flavors present in Jar B: cherry, grape, apple, and lemon. DO: (Show student Card 4) SAY: Suppose I were to reach into the jar without looking and select a single Jolly Rancher candy, look at it, record its flavor, and return it to the jar. How many times would I need to draw out a single candy – replacing it after each draw – in order to be *very confident* about how many candies of each flavor are in Jar B? Explain.

Table C.4. Tasks included in the 4th sequence

Task4PreI	Task6Inst	Task7Inst	Task11PostI
DO: Have the student examine a plastic cup. **SAY:** While setting out plastic cups for a potluck (picnic), you drop several of them on the floor. You notice that some of the cups landed on their side, others landed right side up and still others landed upside down. Suppose you wanted to estimate the likelihood that the cup would land upside down. How many times would you need to drop a cup in order to be confident in your estimate?	There is a lake nearby that has two types of fish in it, Blue Bass and Green Gills. The lake was just stocked with fish but no one knows how many fish are in the lake or the number of Blue Bass and Green Gills. There is a fishing contest coming up soon. The winner will be declared based on who catches the most Blue Bass. Assuming that a contestant catches a fish, the sponsors of the contest need to know the probability of that fish being a Blue Bass. The Probability Explorer has been set up to model the lake containing Blue Bass and Green Gills. You need to simulate catching fish, record and analyze the data, and estimate the probability that a fish is a Blue Bass. Since you can NOT keep fish out of water, every time you catch a fish, you have to throw it back into the lake. Thus, you are doing an experiment *with replacement*. Use Probability Explorer to run **10** trials. Save a copy of either a pie or bar graph in the Notebook. Repeat this 6 times, clearing the data after each set of 10. 1.Based on the data you collected, what would be your estimate of the probability that a fish caught is a Blue Bass? Explain your reasoning. Use Probability Explorer to run **30** trials. Be sure to save an image of either the bar or pie graph in your notebook. Repeat this 6 times, clearing your data after each set of 30. 2.Based on the data you collected for the sets of 30 trials, estimate the probability that a fish caught is a Blue Bass. 3.How did you use your data to make your estimate? 4.Did your estimate change from Part I when you ran sets of 10 trials? Why or why not? Clear all your data. With your partner, use the tools in Probability Explorer to run some more simulations and collect data that would help you make your best estimate of the probability that a fish caught is a Blue Brass.	Your school is planning to create a board game modeled on the classic game of *Monopoly*™. The game is to be called *Schoolopoly* and, like *Monopoly*™, will be played with dice. Because many copies of the game expect to be sold, companies are competing for the contract to supply dice for *Schoolopoly*. Some companies have been accused of making poor quality dice and these are to be avoided since players must believe the dice they are using are actually "fair." Each company has provided a sample die for analysis and you will be assigned one company to investigate. Working with your partner, investigate whether the die sent to you by the company is, in fact, fair. That is, are all six outcomes equally likely to occur? You will need to create a poster to present to the School Board. The following three questions should be answered on your poster: 1. Would you recommend that dice be purchased from the company you investigated? 2. What evidence do you have that the die you tested is fair or unfair? 3. Use your experimental results to estimate the theoretical probability of each outcome, 1-6, of the die you tested. Use Probability Explorer to collect data from simulated rolls of the die. Copy any graphs and screen shots you want to use as evidence and paste them in a Word document. Later, you will be able to print these.	**DO:** Have the student examine a toothpaste cap. **SAY:** While brushing your teeth one morning, you drop the toothpaste cap on the floor. You notice that it landed on its side and figured that it also could have landed right side up or upside down. Suppose you wanted to estimate the likelihood that the toothpaste cap would land upside down. How many times would you need to drop the toothpaste cap in order to be confident in your estimate?