ELEMENTARY PRESERVICE TEACHERS'

CONCEPTIONS OF VARIATION

by

DANIEL LEE CANADA

A dissertation submitted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY
in
MATHEMATICS EDUCATION

Portland State University
2004

TABLE OF CONTENTS

LIST OF TABLES iii

LIST OF FIGURES iv

CHAPTER ONE: Introduction 1

Primacy of Variation 2
Understanding Variation: A Paucity of Research 4
Curricular Inclusion of Stochastics: A Brief History 6
Importance of Teachers' Knowledge 9
Objectives of the Study 13

CHAPTER TWO: Literature Review and Initial Conceptual Framework 15

Foundational Aspects of Probabilistic Thinking 16
Heuristics, Biases, Intuitions, and Misconceptions 18
Aspects of Data Handling: Graphicacy 23
Aspects of Data Handling: Averages 27
Variation in Data Sets 32
Variation in Sampling Situations 37
Variation in Probability Situations 51
Expecting Variation 56
Displaying Variation 58
Interpreting Variation 62
Contexts for Understanding Variation 66
Summary 68

CHAPTER THREE: Methodology 70

Research Design 70
Content and Pedagogy of MET 2 71
Student Characteristics 73
Overview of Research Design 76
Rationale for Design 79
Data Gathering 85
Data Analysis 101
Summary 112

CHAPTER FOUR: Results and Analysis 114

Evolving Framework 114
[1] Expecting Variation 116
[2] Displaying Variation 135
[3] Interpreting Variation 161
Individual Cases 179
The Case of DS 182
The Case of GP 198
The Case of EM 212
The Case of JM 225
The Case of SP 239
The Case of RL 251
Cross-Case Comparisons 268

CHAPTER FIVE: Discussion and Conclusion 278

First Research Question 278
Second Research Question 289
Third Research Question 292
Limitations of Research 301
Implications for Research and Teaching 302
Concluding Comments 306

REFERENCES 307

APPENDICES 324

Appendix A: Informed Consent 324
Appendix B: Surveys and Interviews 326
PreSurvey 327
PreInterview 336
Data & Graphs PostSurvey 348
Sampling PostSurvey 352
Probability PostSurvey 355
PostInterview 357
Appendix C: Class Interventions 369

LIST OF TABLES

Table #	*Table Title*	*Page #*
1	MET 2 Class Profile	74
2	Overall Research Design	77
3	Contexts for PreSurvey (Part 2) Questions	87
4	Contexts for PreInterview Questions	90
5	Summary of Data & Graphs PostSurvey Questions	93
6	Summary of Sampling PostSurvey Questions	95
7	Summary of Probability PostSurvey Questions	98
8	Contexts for PostInterview Questions	100
9	How many times might the arrow land on black? Why?	105
10	How do you think his results on the second set of 50 spins will compare with the results of his first set?	107
11	Write a list to describe what might happen in six sets of 50 spins. Why did you choose those numbers?	110
12	Isomorphism of Interview Questions	181
13	CodeFrame for PreInterview Q2 (Cross-Case Analysis)	269
14	CodeFrame Summary	271
15	Match 6 Rows from PreInterview CodeFrames	272
16	Match 6 Rows from PostInterview CodeFrames	275

LIST OF FIGURES

Figure #	*Figure Title*	*Page #*
1	Height of Four Children	24
2	Number of Raisins in a Box	25
3	Ungrouped & Grouped Data	26
4	Yellow Class & Brown Class	30
5	Minutes to Travel to School	33
6	Initial Conceptual Framework	55
7	Change Brought to Class	59
8	Percentage of Red M&Ms	61
9	Class A & Class B	63
10	Known Mixture Activity	94
11	River Crossing Game	97
12	Initial Conceptual Framework	114
13	Evolving Framework	115
14	GP's Response to PreSurvey Q4	136
15	BP's Response to Data & Graphs PostSurvey Q1c	137
16	JL's Response to Probability PostSurvey Q3	138
17	SP's Response to PreSurvey Q4	139
18	SW's Response to PreSurvey Q4	140
19	MA's Response to Data & Graphs PostSurvey Q1c	141
20	RB's Response to Data & Graphs PostSurvey Q1c	141

21	BP's Response to Probability PostSurvey Q3	142
22	PreInterview Q3 "Graph: 30"	187
23	PostInterview Q3 "Graph: 30"	188
24	PreInterview Q7 "Compare Graphs"	189
25	PostInterview Q7 "Compare Graphs"	190
26	PreInterview Q13 "Likelier Graph?"	191
27	PostInterview Q13 "Likelier Graph?"	192
28	PreInterview Q4 "Graph: 300"	196
29	PostInterview Q4 "Graph: 300"	197
30	PreInterview Q8 "MAX Wait-Times"	207
31	PostInterview Q9 "Muffin Weights"	208
32	Evolving Framework	279

CHAPTER ONE

Introduction

The purpose of this study is to research elementary preservice teachers' conceptions of variation. After defining variation in the introductory section, this chapter discusses four components which motivate the study. The first component is the primacy of variation to statistics and probability (which are together sometimes referred to as stochastics), and the second is the dearth of research in understanding variation. The third component is the inclusion of stochastics in school curricula, and the fourth is the importance of teachers' subject matter knowledge. Taken together, these four components build a case for why elementary preservice teachers' conceptions of variation are a relevant and significant area to investigate in the overall picture of mathematics education. The chapter culminates with a presentation of the objectives of the study and statement of the research questions.

Variation is a term with several related forms and uses. *The Oxford Dictionary of Current English* says that variation refers to a "departure from the normal kind, amount, a standard, etc" (Thompson, 1998, p. 1018). Related in Latin to other terms like *vary, variability,* and *various,* variation implies diversity. A light rail system may promote its trains as arriving at any given stop every ten minutes, but the actual time between arrivals varies. The interval of time is not uniformly ten minutes, and the absence of uniformity indicates the presence of variability. Variation can also refer to the amount of diversity. One way to measure variation in a set of numeric data is to compute the *range*, which gives the spread of the data from the maximum to the

minimum. Another way of measuring variation is to calculate the *variance* and standard *deviation* of the data set, which quantify the spread of the data about the arithmetic mean. The key element to any discussion of variation is that there are differences in the set under study, because without differences there is no variation.

Primacy of Variation

Moore (1990), in a treatise on the nature of statistical thinking, lists five core elements:

1. The omnipresence of variation in processes,
2. The need for data about processes,
3. The design of data production with variation in mind,
4. The quantification of variation,
5. The explanation of variation (p. 135).

Notice that variation is mentioned in four of the five core elements listed by Moore. A report of the joint curriculum committee of the American Statistical Association (ASA) and the Mathematical Association of America (MAA) supports not only the omnipresence of variation as one of their elements of statistical thinking, but also the elements of "measuring and modeling variation" (p. 127). The "*omnipresence of variability*" was cited as giving rise to the very need for the discipline of statistics (Cobb & Moore, 1997, p. 801, italics in original).

The idea that variability is everywhere makes sense when thinking about the world in which we live. Not only do people and their environments vary, but even repeated measurements on the same person or thing can vary (Wild & Pfannkuch, 1999). Also, "natural variation appears in the heights, reading scores, or incomes of a group of people" (Moore, 1990, p. 98). There is also a chance variation component to our world. Moore (1990) points out that one use of probability instruction is to lead

students to the understanding that chance variation, as opposed to deterministic causes, explains most outcomes in our world. He writes:

> It is perhaps surprising that patterns in careful measurements or in data on many individuals can be described by the same mathematics that describes the outcomes of chance devices. Experience with variation is a first step toward recognizing the connection between statistics and probability (p. 99).

Philosophically, living in a stochasticized world implies an existence beset by variation on all sides (Davis & Hersch, 1986); mathematically, "statistics provides means for dealing with data that take into account the omnipresence of variability" (Cobb & Moore, 1997, p. 801).

In addition to the more academic examples cited above, professional statisticians also see the centrality of variation in their work, and others have framed a model of statistical thinking in which variation is the core element (Pfannkuch, 1997; Pfannkuch & Wild, 1998; Wild & Pfannkuch, 1999; Pfannkuch & Wild, 2001). In an investigation of the nature of statistical thinking from the practitioner's perspective, the first component to emerge was that "statistical thinking involves 'noticing', understanding, critically evaluating and distinguishing the types of variation" (Pfannkuch, 1997, p. 407). The theme of variation is pervasive throughout the process of any statistical enquiry. Wild and Pfannkuch (1998) interviewed professional statisticians to capture the stories their subjects wanted to tell. One subject expressed that "basically what distinguishes statistical thinking from anything else is that you accept that variation exists," while another succinctly states that "statistics is the science of variation" (Wild & Pfannkuch, 1998, p. 6). The authors posit that "this very basic element of statistical thinking, 'noticing variation and wondering why', is

actually at the root of much, if not most, scientific research" (p. 7). The mindset of *noticing variation and wondering why*, when coupled with the sense that variability inheres in every facet of life, makes for a rich context in which to embed all other elements of a statistical enquiry. The mandate not only to expect variation, notice variation, but also account for the causes of variation, make up key features of Wild and Pfannkuch's (1999) model of statistical thinking.

The above examples lend credence to the tenet that variation is indeed the central feature behind statistics, and offer support for why others agree that "statisticians consider variation to be the foundation of statistical thinking, the very reason for the existence of their discipline" (Shaughnessy & Ciancetta, 2001).

Understanding Variation: A Paucity of Research

Although variation is central to stochastics, there is relatively little research on people's understanding of this concept. This is not surprising, since as of the late 1980s a review of the literature showed that far more research had been done in the area of probability than in statistics (Garfield & Ahlgren, 1988). Several researchers had attempted to describe stages in the development of probabilistic thinking (Piaget, 1975; Falk, 1983; Fischbein & Gazit, 1983; Green, 1983). Others had focused on intuitive reasoning about probability, revealing not only the kinds of misjudgments people make but also suggesting explanations for these errors (Kahneman & Tversky, 1982; Konold, 1983). Only recently has the attention of researchers turned directly to concept of variation.

In delivering a keynote address to the Mathematics Education Research Group of Australasia (MERGA), Shaughnessy (1997) noted that "although there have been

investigations into students' concepts and beliefs about 'averages', there does not seem to be a similar tradition of research into students' ideas about variability or spread" (p. 5). That address may well have served as a catalyst for researchers to uncover the specific ways in which people thought about variation in different contexts, or at the very least it gave a voice to what other researchers had been noticing as well.

Since that MERGA address, research specifically about conceptions of variation has been slowly emerging, amidst other calls for more research in this area. In a study about data distributions, Mellissinos (1999) comments that "student notions of variability is considered a needed area of research" (p. 1). Concepts of variation, graphicacy, and centers are all factors relating to the research of Watson and Moritz (1999), who remark that "very little research has explored children's strategies involved in comparing data sets" (p. 146). Similarly, Watson and Moritz (2000a) note that little research has been done on students' cognition of sampling situations, for which variation is an integral component. These researchers have consistently called for more research on the understanding of variability in the context of comparing data sets and sampling. The current situation is summarized nicely by Torok and Watson (2000), who wrote that "an appreciation of variation is central to statistical thinking, but very little research has focused directly on students' understanding of variation" (p. 147).

The research which does exist on the concept of variation was mostly aimed at students in grades 3-12. Reading and Shaughnessy (2001) used subjects from grades 4-12, and Watson et. al. (2002) used subjects from grades 3, 5, 7, and 9. However,

while needed research has only recently been and is still being conducted on *school students'* understanding of variation, research has provided few exploratory results on the conceptions of variation held by *teachers*. As others have noted, research on understanding variation is still in its nascency (Watson, Kelly, Callingham, Shaughnessy, 2002; Torok & Watson, 2000; Jones et. al., 2000).

Curricular Inclusion of Stochastics: A Brief History

Stochastics has not always been a vital part of the school curriculum in the U.S. In fact, up until the last decade or so it could safely be said that calls for including stochastics in the American school curriculum had fallen on deaf ears. Despite the encouragement of the 1923 report *The Reorganization of Mathematics in Secondary Education* to include some stochastics among the usual fare of algebra, geometry, and trigonometry, subsequent papers such as the 1938 report *Mathematics in General Education* and the 1940 report *The Place of Mathematics in Secondary Education* placed less emphasis on the importance of stochastics (Bidwell & Clason, 1970). By the era of New Math, the secondary curriculum piloted by the University of Illinois in 1958 comprised eleven units, none of which specifically addressed stochastics (NCTM, 1970). It is therefore no wonder that in 1959, when the College Entrance Examination Board proposed curriculum including a unit introducing probability with statistical applications, they were able to applaud themselves for "one of the more novel suggestions of the Commission" (Bidwell & Clason, 1970, p. 703). Later, the 1963 Cambridge report *Goals for School Mathematics* emphasized an "elementary feeling for probability and statistics" (Cambridge Conference on School Mathematics, 1963, p. 9), but by the time this emphasis was reiterated in the 1975

NACOME report, the actual representation of stochastics in the curriculum was still meager (Gawronski & McLeod, 1980). This brief historical review helps bolster the claim of "the traditional complete absence of stochastics from the school curriculum" (Shaughnessy, 1992, p. 467), and lets us more fully appreciate this situation.

The National Council of Teachers of Mathematics called for increased teaching and learning in stochastics in their 1980 publication *An Agenda for Action*, but it was the NCTM *Standards* of 1989 which gave stature to the place of stochastics among other curricular strands in the United States. (NCTM, 1980; NCTM, 1989). This place has been affirmed in the subsequent release of the *Principles and Standards for School Mathematics* (NCTM, 2000). Also helping drive up interest in stochastics has been the 1997 addition of statistics to the list of Advanced Placement exams. The syllabus for this exam includes data exploration, study design, probability distributions through simulation, and inference. The growth of participation in this exam has been steep. For example, while about 7,600 high school students around the world took the first AP Statistics exam in 1997, that number had risen to over 65,000 by 2004.

However, the rise to prominence of stochastics in the school curriculum is not constrained to the United States alone. Not long before the release of the 1989 *Standards*, Garfield and Ahlgren (1988) cited promising new curricular materials being developed not only in America, but also in the United Kingdom. Furthermore, in reference to developing mathematically literate world citizens, these authors noted the “vigorously growing movement to introduce elements of statistics and probability into the secondary school curriculum, and even the elementary school curriculum” (p. 44). Others affirm the international trend, and point to examples in Spain, Australia,

and New Zealand, in addition to America and the United Kingdom (Batanero, Godino, Valecillos, Green, & Holmes, 1994; Shaughnessy, Garfield, & Greer, 1996; Mellissinos, Ford, & McLeod, 1997; Watson & Moritz, 2000a). In England and Wales, for example, the recommendation for stochastics includes "collecting, representing, and interpreting data" (Department for Education, 1995, p. 10). The curriculum in New Zealand calls for "statistical investigations within a range of meaningful contexts" (Ministry of Education, 1992, p. 186). In reference to notions of sampling and making inferences, *A National Statement on Mathematics for Australian Schools* suggests that "the groundwork should be laid in the early years of schooling in the context of data handling and chance activities" (Australian Education Council, 1991, p. 64). It does indeed seem clear that "topics in data handling have begun to play a more prominent role in the mathematics curricula in many countries" (Shaughnessy et. al., 1996, p. 205).

Looking closer at the PSSM (NCTM, 2000), the following recommendations are made in its Data Analysis and Probability strand:

> Instructional programs from prekindergarten through grade 12 should enable all students to –
> - Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them;
> - Select and use appropriate statistical methods to analyze data;
> - Develop and evaluate inferences and predictions that are based on data;
> - Understand and apply basic concepts of probability (p.48).

It is worth noting that both aspects of stochastics – probability and statistics – are intertwined in the same curricular strand, and this illustrates the way in which concepts of chance are considered integral to a holistic perspective of data handling.

While some may suggest that data analysis take place in an environment wholly separate from probability, the position taken in this study and by the NCTM advocates the synthesis of concepts of data and chance in the school curriculum. Under this position, "in the context of data analysis, or statistics, probability can be thought of as the study of potential patterns in outcomes that have not yet been observed" (Scheaffer, 2002, p. 6). That the full meaning of data analysis should incorporate both aspects of statistics and probability is a perspective shared by many others as well (Jones, Thornton, Langrall, Mooney, Perry, & Putt, 2000; Torok & Watson, 2000; Shaughnessy et. al., 1996).

Importance of Teachers' Knowledge

Since stochastics continues to be emphasized in the school curricula, and since variation is a vital element of stochastics, it makes sense to wonder what teachers know about variation. Quinn (1997) identified stochastics as one of the three "problematic areas of preservice elementary education" (p. 112), along with geometry and rational numbers. Lajoie and Romberg (1998) agree that stochastics may be as new a topic for teachers as for children, and that "teachers must be provided with appropriate preservice and inservice training that will give them the knowledge base they need to feel comfortable teaching about data and chance" (p. xv).

Implicit in stressing the importance of teachers' knowledge is the belief that teachers themselves are important to the enterprise of learning. While this may seem intuitive, it depends on what we mean by "knowledge." The claim of the 1966 Coleman Report was that "teachers, or more accurately variations among teachers, do not make a difference in school achievement" (Shulman, 1988, p. 10). The process -

product research program showed that "teachers did make a difference" (p. 10), but the focus of that program was more on teachers' classroom behavior, not on their knowledge. Much of the past research "showed little or no statistical relationship between teacher knowledge and student achievement" (Grossman, Wilson, & Shulman, 1989, p. 25). One reason for these results is that teachers' knowledge was measured in very limited ways, by looking at the number of classes taken or a teacher's grade point average. This impoverished view suggested that teacher knowledge wasn't vital to student learning, even though this connection may appear to be a matter of common sense (Lockwood, 1998; Fennema & Franke, 1992). In more recent times, a report by the National Commission on Teaching and America's Future cited growing research that "what teachers know and do is one of the most important influences in what students learn" (Darling-Hammond, 1998, p. 6).

Certainly there are many dimensions to the enterprise of teaching, such as teachers' beliefs, knowledge, attitudes, skills, and classroom behavior (Shulman, 1988; Borko et. al., 1992). While teachers may make modifications to their practice along these dimensions throughout their careers, for many of them the initial experience of teacher training provides groundwork in these areas. Foremost among the goals for a teacher education program is that preservice teachers begin to gain the components of knowledge which research suggests are important for teaching (Cooney, 1994). Perhaps because teacher knowledge is an incredibly complex issue (Lehrer & Franke, 1992), it has been studied in terms of different types or components of knowledge. Two components which research has begun to delineate are subject matter knowledge and pedagogical content knowledge.

Shulman (1988) gave a definition of subject matter knowledge as "that comprehension of the subject appropriate to a content specialist in that domain" (p.26), a comprehension which includes "the key facts, concepts, principles, and explanatory framework of a discipline" (Borko et. al., 1992, p.195). The importance of subject matter knowledge is echoed in the *Professional Standards for Teaching Mathematics* (NCTM, 1991), which includes a section addressing the professional development of teachers. The professional development section, Standard 2 entitled "Knowing Mathematics and School Mathematics", states that "teachers of mathematics should develop their knowledge of the content and discourse of mathematics, including mathematical concepts, procedures, and the connections among them..." (p. 132). This description seems very much in line with the idea that subject matter knowledge is about knowledge *of* mathematics and knowledge *about* mathematics (Simon, 1993). Standard 4, "Knowing Mathematics Pedagogy," speaks about "teachers' knowledge and ability to use and evaluate...ways to represent mathematics concepts and procedures..." (NCTM, 1991, p. 151), and this relates to pedagogical content knowledge. This idea of representing mathematics is mentioned by Shulman (1986), who includes in his definition of pedagogical content knowledge the use of analogies, examples, illustrations, and demonstrations, "in a word, the ways of representing and formulating the subject that make it comprehensible to others" (p. 9). Knowing how to accomplish this representation – how to select appropriate tasks, ask good questions, and assess what students understand – is seen as the heart of pedagogical content knowledge (McDiarmid, Ball, & Anderson, 1989; Borko et. al., 1992; Tirosh, 2000).

My study is primarily concerned with the subject matter knowledge of teachers regarding the topic of statistical variation. While pedagogical content knowledge is also important, the research on how children learn about variation is still sparse. Hence, it is difficult to know just how to help teachers work with their pupils when research is still painting the picture of how students think about this topic. It is also seems problematic to address pedagogical content knowledge when even the teachers' own subject matter knowledge is unclear. In the case of elementary preservice teachers, it is unknown what conceptions of variation they have.

Concerning the recommendations of the PSSM (NCTM, 2000) for the Data Analysis and Probability strand for grades 3-5, expectations include: The design of investigations; the consideration of the effects of different methods of data collection; the comparisons of data distributions; and the proposal and justification of predictions and conclusions based on data. Variation is an inherent concept within each of these expectations in probability and statistics, at the elementary level. The research on student thinking, presented in the next chapter, is beginning to inform us of what elementary students can understand about variation. The big question is: What about the subject matter knowledge of the prospective teachers of these schoolchildren? Do preservice teachers participate in experiences where they themselves can develop an understanding and appreciation of variation? If teachers are expected to gain some requisite knowledge at their training institutions, then the colleges and universities should be a place teachers can learn about variation. More than just traditional math courses are called for, and Simon (1993) suggests that " in order to break the cycle of teachers with weak conceptual backgrounds providing conceptually impoverished

instruction, preservice mathematics courses will need to prepare prospective teachers more adequately" (p. 252).

Objectives of the Study

Within stochastics variation is the dominant characteristic of statistical thinking. While research on probability and statistics has produced findings in a number of areas which I'll bring up during the literature review in the next chapter, research specific to the concept of variation has only recently surfaced. Moreover, the research on variation which has come to light has predominantly been conducted with precollege students. Missing from the literature is an idea of how preservice or inservice elementary teachers reason about variation.

Thus, one objective of this study is to develop a framework to characterize the conceptions about variation held by elementary preservice teachers (EPSTs). A second objective is to compare EPTS' conceptions of variation before and after an instructional intervention focusing on variation. A third objective is to investigate types of tasks that might be useful to uncover EPSTs' thinking about variation. My research questions directly reflect these objectives of the study:

1. What are the components of a conceptual framework that help characterize EPSTs' thinking about variation?
2. How do EPSTs' conceptions of variation before an instructional intervention compare to those conceptions after the intervention?
3. What tasks are useful for examining EPSTs' conceptions of variation in the contexts of sampling, data & graphs, and probability?

The next four chapters show how this study unfolded to address the above research questions. Chapter Two includes a description of the previous research on conceptions of variation that has been done in the following three contexts: Variation in data sets, variation in sampling, and variation in chance situations. Also, Chapter Two discusses an initial conceptual framework that was inspired by prior research and a pilot study. Chapter Three provides the methodology used to gather and analyze the data, and also contains a detailed description of how the initial conceptual framework developed into a richer, evolving framework. The meaning of each element in the evolving framework is discussed as part of the results in Chapter Four. Also in Chapter Four, the evolving framework is used to compare EPSTs' conceptions on various tasks from both before and after the instructional interventions. A summary of results is given in Chapter Five, as are implications for teaching and recommendations for future research.

CHAPTER TWO

Literature Review and Initial Conceptual Framework

The first purpose of this chapter is to present a selected review of the research on stochastical teaching and learning that is germane to the study of variation. By situating the study within the field of stochastics education, we can see how previous research in probability and statistics relates to the specific issue of knowledge about variation. The second purpose of this chapter is to articulate an initial conceptual framework that comprises three aspects of expecting, displaying, and interpreting variation. These three aspects were hypothesized as a useful lens for looking at EPSTs' conceptions of variation, and provided an organizational structure for the rest of the study.

Literature Review

In structuring this literature review, first some studies that look at foundational aspects of probabilistic thinking are discussed. Next, examples of research on judgment heuristics, biases, and other stochastical intuitions and misconceptions are provided. Then, some findings on graphicacy and averages that focus on data handling are explicated. I included the above areas because of their connections to reasoning about variation. Lastly, some emergent research on students' concepts of variation in data sets and in sampling and probability contexts is presented. Although this literature review focuses directly on issues of variation, connections are drawn throughout the chapter on how other research relates to and informs the present study.

Some of the research has more of a probabilistic flavor and other research more statistical, but is in some sense an artificial division which separates these twin domains of stochastics. Both domains enhance and influence one another. The philosophical approach of this study is aligned with "forging connections between the study of data analysis and probability concepts" (Shaughnessy, Garfield, and Greer, 1996, p. 206).

Foundational Aspects of Probabilistic Thinking

In choosing this topic for launching the literature review, my main motive is to focus on people's appreciation of random phenomena. The ability to discriminate between the deterministic and the random lies at the heart of stochastics. In particular, the lack of appreciation for randomness can hinder one's ability to deal appropriately with variation.

Piaget and Inhelder (1975) indicated that the age threshold for appreciating the unpredictability of random phenomena is seven years, but Kuzmak and Gelman's (1986) evidence suggested "an earlier understanding of random phenomena than previously has been reported" (p. 565). Fischbein, Pampu, and Minzat (1975) provided evidence that children as young as five have some sensitivity to uncertainty. Kuzmak and Gelman (1986) used both a random device and a deterministic device in an experiment with young children. The deterministic device consisted of a transparent tube which dispensed colored balls one at a time. The color of the next ball could readily be seen. The random device was a rotating wire cage full of colored balls which could dispense these balls one at a time through an attached cup. Children who used these devices were asked whether or they not they knew for sure which

color they were going to obtain. The researchers found that children between the ages of four and seven were able to understand that one cannot say for certain what will be the outcome when dealing with random events. The lack of certainty for individual outcomes is the essential feature of randomness which distinguishes it from determinism.

Research has found that the inclination to cling to a deterministic outlook even in the face of a random process is tenacious. Working with second-graders, Horvath and Lehrer (1998) discuss how some of the children initially "did not think of rolling dice as completely random", but that "beliefs about lucky numbers and partial telekinesis usually did not endure long after the children's first opportunities to collect data" (p. 126). In a questionnaire item administered to 1014 students from grades 3, 6, and 9, 113 responses affirmed some kind of belief in lucky numbers, as exemplified by the student who wrote: "I don't think many numbers are lucky. But I think 4, 7, and 9 are, so I guess I'd agree in a way you can have lucky numbers" (Watson, Collis, and Moritz, 1995, p. 553). This attitude, similarly exemplified by the student cited by Horvath and Lehrer (1998) who said "I usually roll 6s" (p. 137), begs for an examination of what is really involved in a random event.

An appreciation for randomness cannot be overemphasized as an important facet for understanding variation, and represents something "fundamental to reasoning within the domain of statistics and probability" (Metz, 1998, p. 156). Determinism thwarts randomness, and thus stifles an appreciation of variation. Moreover, it would

be an erroneous assumption to assume that by the time people reach a certain age, they have completely abandoned deterministic notions and embraced the truth of living in a stochasticized world.

Heuristics, Biases, Intuitions and Misconceptions

A host of misconceptions related to probabilistic and statistical thinking have been researched, and psychologists in particular have characterized many examples of faulty human intuition in the face of uncertainty. The use of the term *misconception* in this area of research refers primarily to erroneous thinking which contrasts with "correct", normative responses predicted by statistical theory. This section illustrates some of the key findings that research on intuitive stochastical thinking has produced. Because this domain of research is so robust and stable, it suggests that when subjects reason about variation, they will also bring with them strong intuitions that influence their thinking.

Psychologists Daniel Kahneman and Amos Tversky found that people come to the table of stochastical learning with a host of their own intuitions about the subject, and that these intuitions often serve them quite poorly. As they wrote early on, "We submit that people view a sample randomly drawn from a population as highly representative, that is, similar to the population in all essential characteristics" (Kahneman & Tversky, 1971, p. 105). Some examples will illustrate the key features of this judgment heuristic, which is known as *representativeness*.

For the first example, in considering the gender of six children in a family, many people consider the sequence of GGBGBB to be more likely to occur than BBBBBG. Many people think that any sample drawn from the population should

reflect a near 50-50 distribution of boys and girls (Kahneman & Tversky, 1972; Shaughnessy, 1977). Kahneman and Tversky also identified what they called the "law of small numbers, which asserts that the law of large numbers applies to small numbers as well" (1971, p. 106). Shrage (1983) also observed this in his tertiary students. Using a population mixture of 50% white balls and 50% black balls, he found that students believed the chances of getting 7 white balls in 10 draws and the chances of getting 70 white balls in 100 draws were the same. People often intuitively rely on the law of small numbers when playing games of chance.

If the results of a sequence stray too far from the population proportion, a "corrective bias in the other direction is expected" (Kahneman & Tversky, 1971, p. 106). This helps explain why a person who sees flips of a fair coin result in five heads in a row will intuitively think there is a higher likelihood of a tails on the sixth flip, a psychological phenomena known as the *Gambler's Fallacy*.

The representativeness heuristic also suggests that any uncertain event must "also reflect the properties of the uncertain process by which it is generated" (Kahneman & Tversky, 1972, p. 434). For example, when flipping a coin, the representativeness heuristic implies that any sequence of flips should appear random. Thus, people do not consider sequences such as THTHTHTHTH or TTTTTHHHHH to be representative, although both have a 50-50 mix of Hs and Ts. These sequences appear to be too ordered. They do not appear to be random, and thus do not "represent the fairness of the coin" (Tversky & Kahneman, 1974, p. 1125).

A second judgment heuristic, called *availability*, is used when people ascribe likelihoods to situations based on how readily they bring similar examples to mind (Tversky & Kahneman, 1974). An example of this bias occurs when people are asked if it is easier to choose committees of two or committees of eight from a group of ten. Those operating with an availability heuristic will find it easier to think of examples of committees of two, and hence they fail to recognize the equivalent nature of the problem (Shaughnessy, 1993). Similarly, people erroneously tend to think there are more words beginning with "r" than there are words with "r" as the third letter, because words starting with "r" are easier to recall (Tversky & Kahneman, 1974). Reliance on the availability heuristic leads people to depend on what they can recall or mentally construct most readily.

Another facet of intuitive thinking discovered in research is the *outcome approach*, which colors a person's fundamental understanding of the goal of probability. People who operate under the outcome approach interpret probability questions as "predicting the results of a single trial" (Konold, Pollatsek, Well, Lohmeier, & Lipson, 1993, p. 394). For example, consider a six-sided die with five black sides and one white side. One student reasoned that since a single toss would "almost certainly" result in black, then by extension six tosses would result in six blacks (Konold, 1989, p. 83). Another component of the outcome approach includes the way in which probabilities are judged as right or wrong after each outcome. For example, Konold (1989) found that, if a weather forecaster predicts a 70% chance of rain for tomorrow, and then it doesn't rain, some subjects evaluate the 70% prediction

as erroneous. That is, the forecaster's prediction is judged wrong because the predicted outcome *didn't* occur. Outcome approach thinking can also result in an emphasis on causal features rather than frequency data. Some people in Konold's (1989) study inspected the features of a seven-sided irregular polyhedron, or the way in which it was rolled, to predict outcomes, and these people discounted the actual results of 1,000 previous rolls. A view towards an equiprobable interpretation may interact with outcome approach thinking. A person may view all outcomes of an event as equally likely, and deem the goal of probability questions as simply a matter of choosing any outcome which *could* occur (Shaughnessy & Ciancetta, 2001; Shaughnessy & Bergman, 1993; Shaughnessy, 2001).

These three types of intuitive stochastical thinking (representativeness, availability, and the outcome approach) are examples of cognitive strategies which may influence a person's reasoning in situations where attention to variation is critical. For instance, representativeness leads to a person's insensitivity to the effects of sample size. This is readily illustrated in the results shown for what is known as the Hospital Problem (Tversky & Kahneman, 1974), which is summarized as follows:

> A certain town has two hospitals. In the larger hospital, about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. For a period of 1 year, each hospital recorded the days on which more than 60% of the births were boys. Which hospital do you think recorded more such days?

The researchers noted that 53% of undergraduate students thought that the two hospitals would have recorded about the same number of such days. The subjects did not recognize that the larger hospital would be less likely to deviate from the expected 50-50 gender mix, and "This fundamental notion of statistics is evidently not part of

people's repertoire of intuitions" (1974, p. 1125). Availability may lead a person to attach inflated significance to an event which has personal meaning. Such an event becomes more than just another data point, and can bias statistical thinking. Shaughnessy (1992) offers this example, "If several of your friends have recently divorced their spouses, you may be led to believe that the local incidence of divorce is on the rise, when in fact it has not changed" (p. 472). In this sense, availability can reduce the perception of variation. If a child recalls a game of chance in which she rolled many sixes, then it may make sense for her to think that six is a lucky number. The outcome approach also can diminish the perception of variation, as demonstrated by the subjects in Konold's (1989) study who predicted all blacks in the tosses of the die with five black faces and one white face.

To summarize, when considering conceptions of variation it is important to remember the types of judgments that people are susceptible to, and that people "already have their own built-in heuristics, biases, and beliefs about probability and statistics" (Shaughnessy, 1992, p. 472). It is also important to attend to a person's fundamental understanding of randomness. The next two sections discuss research on graphicacy and averages, two aspects of data handling which both influence conceptions of variation. For instance, graphs summarize data and reveal or disguise variation in the data set depending on the type of graph used. An average might be representative of the data set, but variation can address how the data clusters or spreads out around the average.

Aspects of Data Handling: Graphicacy

Graphicacy is a term first introduced by Balchin and Coleman in 1965, and has evolved in definition until Wainer narrowed it to mean "the ability to read graphs, defining it as proficiency in understanding quantitative phenomena that are presented in a graphical way" (Friel & Bright, 1996, p. 1). The facility to construct and interpret graphs begs the question of what exactly constitutes a graph. For the purposes of this study it seems natural to consider the standard sorts of graphs and plots of "univariate data that dominate the school curriculum, that is, line plots, bar graphs, stem-and-leaf plots, and histograms" (p.1). Also, the kinds of pictographs commonly found in the media are of interest in this study, as are rudimentary presentations of bivariate data such as scatter plots or line graphs. This study adopts the position that "the use of graphs and other kinds of representations needs to be viewed as part of the process of statistical investigation and not as an end in itself" (Friel, Bright, Frierson, & Kader, 1997, p. 62). Graphs can be viewed as an important part of data handling, whose chief role lies in data reduction.

Curcio (1987) conducted a study of 204 fourth-grade and 185 seventh-grade subjects. Students were given a test which used twelve graphs equally distributed among the following four types: bar graphs, circle graphs, line graphs, and pictographs. Curcio was able to identify three levels of difficulty in making sense of graphs. The first level is simply *reading the data*, in which the subjects could attend to the basic facts stated in the graph, including numerical values shown, titles, and axis labels. For example, in a bar chart showing the heights of four children (see Figure 1),

a question invoking a simple reading of the data would be to ask, "How tall is Mark?" The second level of comprehension is *reading between the data*, which necessitates "comparisons and the use of mathematical concepts and skills" (Curcio, 1987, p. 384). A question for this level would be, for example, "How much taller is Dan than Mark?" The third level is *reading beyond the data*, which requires "extension, prediction, or inference" (p. 384). To establish the level of statistical literacy essential to functioning in a modern society constantly assailed by data, people need to move up through these hierarchical levels of difficulty of graphical comprehension (Moritz & Watson, 1997; Curcio, 1987).

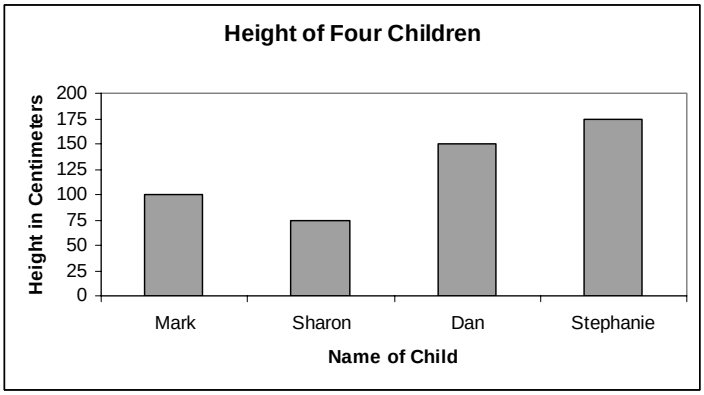


Figure 1 – Height of Four Children

In 1994, Friel and Bright (1996) studied graphicacy among students in grades 6, 7, and 8. They researched the levels of graph comprehension outlined earlier by Curcio (1987). For example, they showed students a line plot depicting the quantity of raisins in half-ounce boxes (see Figure 2). Rather than initially ask students questions about *reading the data*, Friel and Bright started with the following questions aimed at *reading between the data*: "Are there the same number of raisins in each box? How can you tell?" (Friel & Bright, 1996, p. 4).

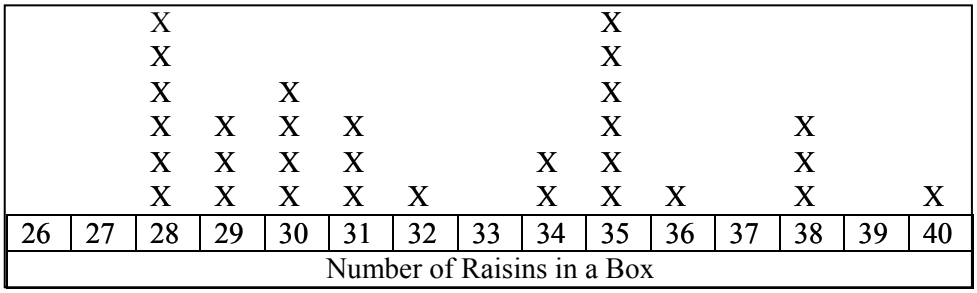


Figure 2 – Number of Raisins in a Box

Results showed that although many students were able to answer the first question correctly in the negative, in their subsequent explanations it became clear that they were looking at the graph in ways that did not support their answer. For example, one student said "No, because they weigh the boxes until they equal ½ ounce. They don't count the raisins" (p. 5).

Bright and Friel (1998) also tested students before and after an instructional unit which was designed to "highlight connections between pairs of graphs" (p. 68). For each pair of graphs, the same set of data was used. Two pairs of graphical types were stem-and-leaf plots versus histograms and line plots versus bar graphs. A third pair was bar graphs for grouped versus ungrouped data (see Figure 3). The graph at the top of Figure 3 represents *ungrouped* data (raw data), and students can simply point to a particular bar to identify their own data value. The graph at the bottom of Figure 3 represents *grouped* data (reduced data), and specific data values are not obtainable for any individual student. Bright and Friel wanted students to determine "not only what information is presented in each representation but also whether identical information could be extracted from the different representations" (1998, p.

69). The researchers found that students had difficulty in making the translation from graphs of ungrouped to grouped data.

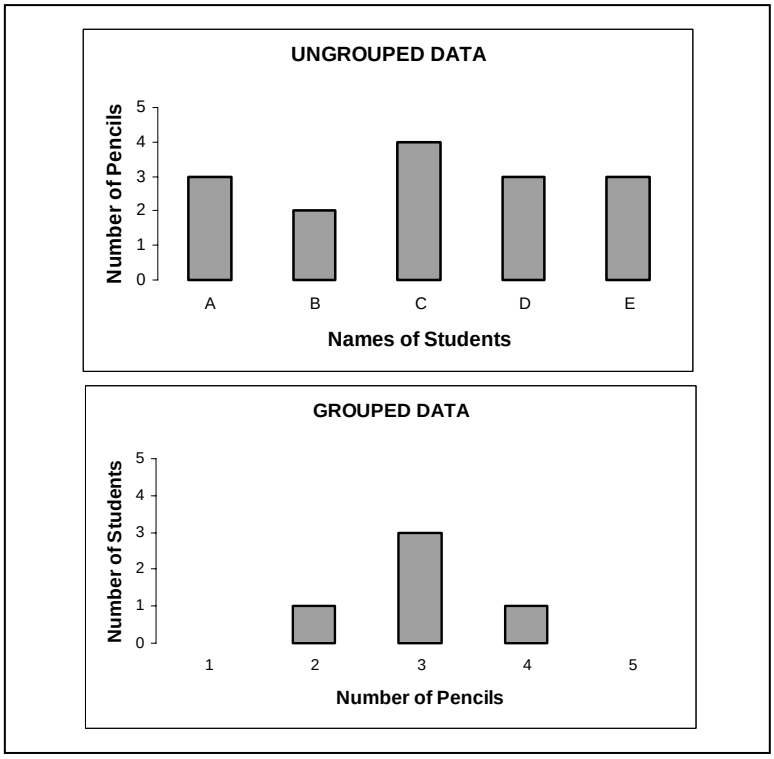


Figure 3 – Ungrouped & Grouped Data

As one student commented when looking at a graph for grouped data, "I don't know how you add this thing up" (p. 74). The central theme propounded by Bright and Friel is that "establishing connections or translations among representations is critical for developing understanding" (p. 82). While their analysis of results does discuss aspects of mode, median, and mean in the students' search for centers, Bright and Friel paid less attention to the ways in which variation reveals itself in the various graphical representations. The lack of attention to subjects' conceptions of variation is noteworthy because many of the researchers' graphs, as well as questions about what

is typical for the data set, seem to provide a natural context in which to look at variability in data presented visually.

In summary, research indicates that there are different levels of graphical understanding held by students, and that the type of graph makes a difference on what people comprehend. It may well be that the research on graphicacy represents a missed opportunity to look at variation in the context of visual data displays (Shaughnessy, 1997), but it can now be seen that a study about variation, if incorporating aspects of graphicacy, can benefit from consideration of what a person can comprehend from graphs.

Aspects of Data Handling: Averages

The concept of average is fundamental to statistical thinking. (Davis & Hersh, 1986). Watson and Moritz (2000c) point out that for a long time the school curriculum focused almost exclusively on the arithmetic mean as being synonymous with finding an average. The notion of average as representative of the data, including the use of measures such as mean, median, and mode, has only been emphasized in the American school curriculum in the past decade. Note that when talking about measures of central tendency, the term *representativeness* is used not as it was in the judgment heuristics discussed earlier. Kahneman and Tversky used the term as a label for a type of erroneous intuitive thinking, while in the context of this section the word has a positive connotation for statistical thinking. Representativeness as it will be used in this context refers to a way of describing how well an average summarizes a set of data.

Work done in a study by Mokros and Russell (1995) of fourth, sixth, and eighth-graders showed that "students' notions of representativeness or typicality grow out of their everyday experiences and have a strong flavor of reasonableness and practicality" (p. 21). Students showed what they knew about averages by constructing different data sets that could reflect a given average. Some of these construction problems were made richer through added constraints, such as the prohibition of using the actual average in the data set, or the mandatory use of preexisting data values.

For example, in the "Potato Chips" problem, the task was to put price tags on 9 bags of chips so that the typical, or average, price would turn out to be $ 1.38. An added constraint to extend the problem was to disallow any price tag from actually being $ 1.38. In the "Allowance Construction" problem, the aim was to construct a distribution of allowances which, taken together, had an average of $1.50. An extension was to specify that 2 allowances had to be 75 cents, and 3 had to be one dollar. Thus, "the student's task was to create a large frequency distribution where some data was already placed" (Mokros, Russell, Weinberg, & Goldsmith, 1990, p. 4). Other problems asked students to interpret what was "typical" from graphs. In the "Allowance Interpretation" problem, a skewed, bimodal graph showing a number of students' allowances was presented. The task was to "use the data to determine the typical allowance as well as the highest amount that could be argued for" (Mokros & Russell, 1995, p. 24). A third type of problem involved understanding a weighted average. The "Elevator Problem" essentially aimed at asking for the average weight of a group of ten people, comprising six men whose average weight was 180 pounds,

and four women whose average weight was 125 pounds (Mokros et. al., 1990; Mokros & Russell, 1995).

Of the five predominant approaches to average identified by Mokros and Russell, two approaches - average as mode, and average as algorithm - were not associated with the notion of representativeness. Modal thinkers were easily able to construct data sets, since they saw average as the value occurring most frequently. However, "when they were not allowed to use the average value as part of their distributions, real difficulties were encountered" (Mokros et. al., 1990, p. 7). In general, students found it much harder to work from the average to the data than vice versa. The standard algorithm for finding an average doesn't go far in helping to solve construction problems, as is exemplified by one student whose first attempt at solving the "Potato Chip" problem was to take the desired average of $ 1.38, multiply that by 9, and then divide by 9.

> When the interviewer asked if there's any way she could put some prices on the chip bags, she replied that she knew how to get an average, but had not yet learned how to find the "numbers that go into an average" (p.15).

The other three approaches - average as reasonable, or as midpoint, or as a point of balance - exemplified a sounder understanding of what it means for an average to represent a set of data. Average as "reasonable" was considered by Mokros and Russell to be significant for a more robust understanding about the concept.

When considering a data set, two fourth graders and two sixth graders demonstrated the key features of what the researchers mean about the notion of average as reasonable. First, the students relied on values that made sense in the

context of the problem, in line with their own understanding of prices, allowances, or weights. Second, the students had some regard for the idea of average as roughly centered rather than precisely in the middle of the data, and that in some sense "high values must be countered by low values" (p. 9). Another important aspect of the research was the finding of how strong an attraction symmetry had for some students, particularly those who gravitated towards a midpoint strategy (Mokros & Russell, 1995). The researchers concluded that premature introduction of the algorithm for finding the mean may in fact be impede a students' overall understanding of the ways in which the mean is or is not representative of the data.

To broaden children's' narrow view of average, Mokros and Russell suggested it is important to "focus on describing and comparing data sets" (1995, p. 37). This suggestion was central to a study of eighty-eight students in grades 3 to 9 on the emergent ideas of statistical inference by Watson and Moritz (1999), as was the notion of average as a representative measure.

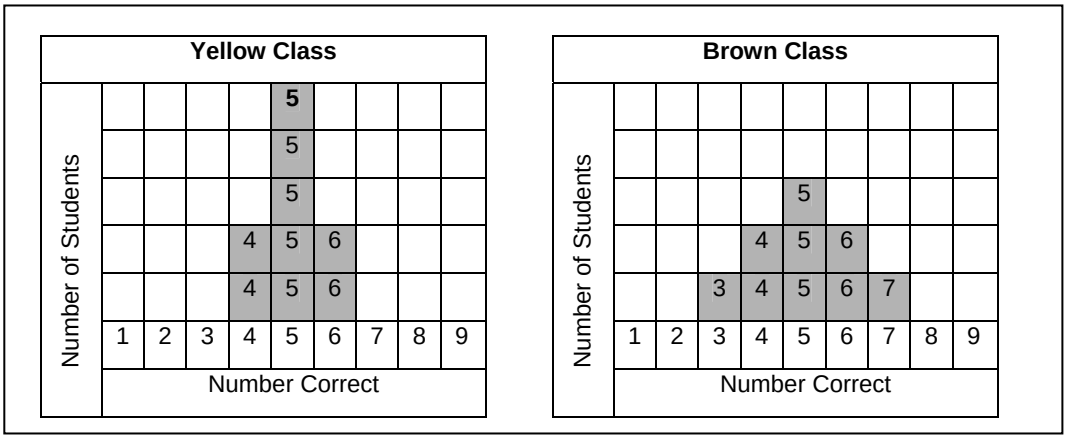


Figure 4 – Yellow Class & Brown Class

The researchers asked students to compare the performances of the two classes whose

test scores were shown in frequency graphs (see Figure 4). Students were asked which class did "better". The researchers classified student responses according to two strategies (numerical or visual) which were used either singly or together. For example, some students using a numerical strategy saw the graphs as a way to obtain the actual scores to calculate totals and means for comparing the two classes. Students using visual strategies commented on aspects like the symmetry or spread of the graphs. Some students "commented on both visual and numerical strategies, but expressed conflict between them rather than viewing them as complementary related strategies" (Watson & Moritz, 1999, p. 155).

In the example shown in Figure 4, the two classes were of equal size. Another similar example used classes of unequal size, with the result that proportional reasoning became increasingly important in analyzing the problem. The researchers concluded that students need more experience with a variety of data sets, and with tasks that allow for the representation of data graphically, and more experience with summarizing data with measures of central tendency.

As we turn to look at some studies which synthesize different notions of statistical thinking, it becomes clear that a study on variation cannot properly be divorced from other key ideas such as average and graphicacy. A person's facility with graphs, measures of center, and concept of distribution all may influence how one thinks about variation. In the next section, shades of these different but related notions can be seen, as attention is now focused on research which highlights the contexts in which variation was explicitly revealed or explored. The research on variation now

unfolds in three broad contexts: data sets, sampling, and probability situations. Much of the research which follows exhibits a blend of statistical concepts, such as graphicacy, averages, or distributions. In a way, many of these studies drew on statistical skills, concepts, and intuitions related to those discussed earlier. The main difference is that in many cases variation was the focus of the investigation itself.

Variation in Data Sets

Jones et al. (2000) claim that "for students to exhibit statistical thinking, there is a need for them to understand data-handling concepts that are multi-faceted and develop over time" (p. 271). Data handling incorporates "organizing, describing, representing, and analysing data, with a heavy reliance on visual displays such as diagrams, graphs, charts, and plots" (Shaughnessy, Garfield, & Greer, 1996, p. 205). The kind of thinking implied by these statements includes attention not only to graphicacy and averages, but to spread, or variation, as well. Data handling implies finding ways to reduce the data while retaining the key features of the data set.

Friel et. al. (1997) mention the process of data reduction and the structure of graphs as factors influencing graph knowledge. They note that "data reduction is an essential part of analysis of the data; different graphs emphasize different degrees of data reduction" (p. 62). In assessing the components of data reduction, they used a problem that involved a stem-and-leaf plot (see Figure 5) of minutes taken by middle grade students to get to school. The accompanying question was, "What is the typical time it takes for students to travel to school?" (Friel & Bright, 1996, p.6). The researchers concluded that "students were less likely to compute measures of center as

part of their responses" (Friel et. al., 1997, p. 60).

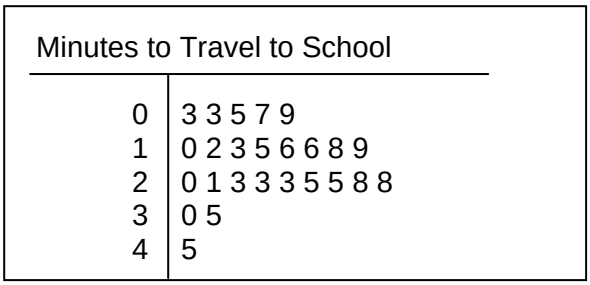


Figure 5 – Minutes to Travel to School

Evaluating the different responses, the researchers seem to validate the idea that attention to variation was an important part of an overall analysis of the question. They note that only a few students chose to use the mean or median in their responses. Instead, results showed that students responded in terms of clusters of typical times, or in terms of a range of numbers that occurred more often, or in terms of the mode. They rhetorically ask of the various methods,

> Is one 'more appropriate' than another; do we want students to move beyond the use of the mode as a tool in this case to using clusters of data as a way of describing what's typical? If so, what is a 'good sized' cluster to be highlighting? (p. 60).

Thus, in the context of the graphical aspects of data reduction, students need an awareness of the importance of both measures of central tendency *and* the spread of the data.

Another approach to looking at students' statistical thinking focused on the concept of distribution. Mellissinos, Ford, and McLeod (1997) noted that although previous studies identified some ideas on how students reasoned about the concept of average, "they have revealed little about how students make sense of an average in the

context of the distribution that it represents or summarizes" (p. 176). Using the "Potato Chip" problem of Mokros and Russell (1995) described earlier, Mellissinos et. al. interviewed a middle-school student. Their results supported the findings by Mokros and Russell (1995), in which the subject did not understand what it meant for an average to be representative of a set of data. What distinguishes the work of Mellissinos et. al. (1997) is their focus on an understanding of distribution. They write, "One reason for the student difficulties may be that students have not learned to think about the mean as a representative measure of a distribution" (p. 179). Mokros and Russell's notion of average as a representative measure involves capturing a range and distribution of a set of data, but Mellissinos et. al. caution that an understanding of the distinction between the terms "distribution" and "data set" is necessary to establish and interpret representativeness. According to Mellissinos et. al. (1997),

> A data set is a collection of measurements of one or more characteristics (of objects or people). A distribution is an attribute of a data set that communicates how measurements in a data set are distributed across its range of values (p. 179).

Mellissinos et. al. conclude that "without a clear idea about distribution, it is difficult to make inferences about student notions of representativeness" (1997, p. 179).

Mellissinos also researched how the notions of a distribution's shape, center, and spread interact (Mellissinos-Lernhardt, 1999). This time she used the "Potato Chip" problem with college students and another task that involved the pulse rates of a set of 30 people in the same age group. One task used both a table and a histogram to present the 30 pulse rates. After first being told that the typical pulse rate for the given

age group was around 65 to 69 beats per minute, students were asked if they thought that the pulse rates for a given set of people were typical. Results showed that, for some students, the ability to interpret a mean did require some concept of the range of possible values. In the Potato Chip task, one student expected variability in the prices for bags of chips, but did not "have a sense of how much variability would be possible for the situation" (Mellissinos-Lernhardt, 1999, p. 6). In the Pulse Rate task, the same student "relies too heavily on the mean to decide whether the group is typical. She does not take into consideration how spread out the pulse rates are, how the data cluster and whether there are any outliers" (p.7).

It thus did seem that some students had some awareness that only looking at the center does not capture the whole picture. Mellissinos reiterates that while many educators promote the mean as representative of a distribution "the concept of distribution relies heavily on the notion of variability, or spread" (p. 1). Certainly the characteristics of a distribution can rely on representing data through summary statistics, but graphs also provide a representation. Hence, Mellissinos' research highlights the connections in statistical thinking between conceptions of graphicacy, centers, and variation, through the common unifying theme of distribution.

Shaughnessy (1997) discusses the work of middle school students on data sets stemming from the weather pages of a local newspaper. Students were gathered in groups of four to five, and each student was given a days' weather page so that their small group had a small set of consecutive days' worth of data on the weather. He observed and noted the types of questions that students came up with, and the approaches they took in analyzing their own questions. In their analysis students

became aware of the kinds of variation inherent in weather data. For example, Shaughnessy wrote that "quite a discussion ensued as to why there was such variability for the coldest time of day" (1997, p. 12). The theme of weather was also used by Torok and Watson (2000), who explored concepts of variation with sixteen students, four each from grades 4, 6, 8, and 10. The two researchers framed their weather task using the concept of average. Students were told that the average maximum daily temperature for Hobart, Australia, over the year was 17 degrees Celsius. Questions included whether or not students thought all the days of the year had that maximum temperature of 17 degrees, and what the maximum temperature might be for 6 different days of the year. They also asked students for likely ranges of temperatures, such as what the highest and lowest maximum daily temperature might be for the month of January, July, or over the course of the year. The researchers found that older students had, in general, a higher level of understanding of variation than the younger students.

Shaughnessy and Pfannkuch (2002) have found that the data sets for the Old Faithful geyser provide an excellent context for highlighting the role of variation in statistical analysis. They describe a classroom exploration in which students were first given one day's worth of data for the number of minutes between successive eruptions of the geyser. The question which threaded through this investigation was about how long one should expect to wait between eruptions of Old Faithful. Students were then told to represent this data in a graph of their choice. Then they were given several more days' worth of data and asked to graph it as well. Some of the students used

boxplots, which do demonstrate the range of the data but which mask the nature of variation which can be seen when the data is plotted over time. Other students used histograms, which reveal the underlying distribution. At first, many students just make an initial prediction based on measures of central tendency (such as the mean or median), which also disregard the variability in the distribution. Shaughnessy and Pfannkuch (2002) point out that

> Students who attend to the variability in the data are much more likely to predict a range of outcomes or an interval for the wait time for Old Faithful, such as "Most of the time you'll wait between 50 to 90 minutes," rather than a single value of 70 minutes (p. 5).

Also, when students first look at one day's data, and then look at several days' data, they may see the variation across days as well as the variation within a day. This extends Curcio's (1987) analysis to what these researchers call looking "behind the data" (p. 6).

The point of the research in this section is that questions about data sets can indeed be shaped to explore student understanding of variation. The trend in the research discussed so far is to recognize the importance of blending of variation with several other statistical concepts such as graphicacy, distributions, and averages. In addition to the contexts of data sets, students' conceptions of variation can also be studied within the context of sampling, which is next discussed.

Variation in Sampling Situations

As mentioned in the previous chapter, the term *variation* and its different linguistic forms come with many related meanings. In sampling situations, variation appears in the differences among repeated samples drawn from the same population.

Samples also vary in the degree to which they represent their parent population. For example, if the population is all students at a high school, using the same opinion poll on two different groups of students is likely to produce two different poll results. Thus, sampling situations can invoke many levels of meaning when we consider variation.

Within the context of sampling, "as sample size increases, the statistics of a sample become less variable and more closely estimate the corresponding parameters of the population from which the sample was selected" (Well, Pollatsek, & Boyce, 1990, p. 289). This is due to the Law of Large Numbers, yet Kahneman and Tversky (1972) found, through tasks like the Hospital Problem mentioned earlier, that many people are unaware of how sample size influences variability. Well et. al. (1990) point out the objections of other researchers, who criticized some of Kahneman and Tversky's problems as being too difficult for subjects to fully comprehend.

Some research has shown that people can be correctly influenced by sample size. For example, in a question that asked which of two samples of different sizes would better estimate some characteristic of a population, Bar-Hillel (1982) found that over 80% of subjects correctly chose the bigger sample. To gain further insight into how well people understood the effects of sample size on the variability of the mean of the sampling distribution, Well et. al. (1990) posed a series of questions to undergraduates,

> in some cases asking subjects to judge which of two samples was more likely to fall closer to the population mean and in others, asking them to judge which of two samples was likely to deviate more from the population mean (p. 292).

One context used by Well et. al. was the average height of American males. Students were told that the national average height of 18-year-old males is 5 feet, 9 inches. They were also told that at Post Office A, 25 men registered for the draft each day for a year, while the number of men registering at Post Office B was 100 men per day for a year. At each Post Office, the average heights of men per day was computed. As an example of a question pertaining to the tails of the distribution, students were asked which Post Office would have recorded more days on which the average height was 6 feet or more. The corresponding question pertaining to center of the distribution was identical, except that "6 feet or more" was replaced by "between 5 feet 6 inches and 6 feet" (Well et. al., 1990, p. 297). Results indicated that people used information on sample size more accurately when dealing with centers of distributions rather than the tails. However, even when the subjects had received instruction on sampling distribution, "many of them still did not understand how sample size influenced the variability of the sample mean" (p. 310). This research does highlight the importance of understanding a distribution, a point also brought out in the research of Mellissinos (1999).

Some of the questions used by Well et. al. are similar in spirit to the Hospital Problem and to a question adopted by Watson, Collis, and Moritz (1995). These latter researchers interviewed a subset of twelve students from 171 girl subjects from grades 3, 5, 7, and 9. They intended to explore the general notion that small samples are more likely to have extreme results than large samples, with this question: "The researchers took a random sample from each school: 50 children from the city school,

20 children from the country school. One of these samples was unusual: it had more than 80% boys" (p. 6). Students were then asked whether the unusual sample was more likely to have come from the city or country school. Not one of the twelve students was able to give a suitable justification for their response. The same question was also used by Watson and Moritz (2000b), who asked 41 students from grades 3, 6, and 9. They found that only six students could give an adequate explanation that connected unusual results to the smaller sample. Many of the other responses made it seem as though the choice of large or small sample was almost a random decision, "with the reasons given for choosing the small samples also being given for choosing the large one" (Watson, 2000a, p. 122). The original Hospital Problem can be theoretically modeled by a binomial distribution, while this modified version above involves a hypergeometric distribution (since children are chosen without replacement). In either case, many students do not recognize that "the smaller sample is more likely to give an extreme or biased result" (Watson & Moritz, 2000b, p. 66).

Fischbein and Schnarch (1997) included the original Hospital Problem in their study on the evolution of probabilistic, intuitively based misconceptions. Of eighty students in grades 5, 7, 9, and 11, plus 18 prospective teachers, only one ninth grader suggested that the smaller hospital would have the more extreme result. In addition, the percentage of respondents who thought the results would be about the same for both hospitals actually increased with age, from 10% of the fifth graders all the way up to 89% of the prospective teachers. Related to the heuristic of representativeness mentioned earlier, the basic misconception here is that sample size doesn't affect

variation. The researchers note that "this misconception developed with age in a surprisingly regular manner" (Fischbein & Schnarch, 1997, p. 101), meaning that more people committed the misjudgment as the age of the subjects increased.

Watson (2000a) gave 33 preservice secondary mathematics teachers the Hospital Problem, and categorized her results by whether the respondents based their solutions solely on intuition, on a mathematical argument, or on a mixture of approaches. She found the success rate of 55%, while not surprising in comparison to results of Kahneman and Tversky's (1972) study of tertiary students, was not related to her subjects' prior formal mathematics experience. Moreover, she notes that "it is disappointing that so few naturally mixed intuition with an attempted mathematical justification in solving the hospital problem" (Watson, 2000a, p. 134).

The Hospital Problem and similar questions not only illustrate the kinds of intuitions and use of heuristics shown by Kahneman and Tversky (1971, 1972), but also allow researchers to investigate students' understanding of the relationship of a sample to its underlying population. As Watson and Moritz put it, such questions allow researchers to "investigate students' understanding of the effect of increased sample size in increasing the reliability with which the sample represents the populations" (2000b, p. 50). Watson and Moritz (2000a) wished to determine if "students recognize the tension between efficiency of small sample sizes, and the reliability of larger sample sizes" (p. 6). They asked 2040 students from grades 3 through 11 whether someone should choose to buy one car in favor of another based on the opinions of a few friends, on the results from a statistical report on 400 cars of

each type, or whether both sources of information were equally valid. More students chose the response that both sources were equally valid over any other response, although the older students more successfully identified the greater reliability inherent in the larger sample. The researchers (Watson & Moritz, 2000a) noted that "some students believe that any sample however small is representative while others believe that 'larger' is always 'better' to achieve representativeness, without regard to increasing difficulty and cost of data collection" (p. 31) . This tension is further revealed in some of the subsequent research on sampling which was undertaken in the context of conducting surveys, or polls.

Jacobs (1997) conducted a study using fourth- and fifth-grade subjects, in which childrens' informal understandings about sampling issues were investigated. Specifically, Jacobs was concerned about students' perceptions of sampling methods, and in students' distinctions between those methods that led to representative samples versus those that led to unrepresentative samples. Some of Jacobs' questions were in the context of taking a survey to predict how many schoolchildren will purchase a ticket for the school raffle. Other questions were in the context of taking a survey to determine how many of the city's schools were recycling. In each context, subjects were presented with a variety of sampling methods to evaluate in conducting the survey, such as simple and stratified random sampling methods, self-selected methods, and restricted methods (which used select groups of the population who would be more likely to skew the results in a certain direction).

Jacobs found that children evaluated these sampling methods by focusing on

the potential for bias, fairness, practical issues, or the results produced by the method. For example, "some children…were able to identify potential bias with restricted and self-selected sampling methods and to recognize a lack of potential bias with random sampling methods" (Jacobs, 1997, p. 12). In focusing on fairness, however, some students were not impressed by the simple random methods which ensured that everyone had the same chance of participating in the survey. They were "not thinking of fair in the probabilistic sense but rather in the affective sense of how the participants (or non-participants) felt about having the opportunity to participate in the survey" (p.12). This interpretation of fairness led subjects to want members from all types of subgroups in the sample, which meant that stratified random sampling was a favored method for these subjects. Some of Jacobs' subjects also focused on the possibility of extreme outcomes, even though the chance of those outcomes occurring was low. Other researchers have claimed that the focus on extreme outcome is suggestive of the outcome approach (Shaughnessy, Watson, Moritz, & Reading, 1999), because even though an outcome might have a small probability, the outcome still *could* occur. Lastly, students assumed a correspondence between the results produced by the sampling method and the results expected prior to sampling. That is, "if the results corresponded with what was expected, then it was an appropriate sampling method because it got the 'right results' " (Jacobs, 1997, p. 14). The converse also held, so that sampling methods which produced different results from what was expected are judged to be incorrect methods.

Statistical inference involves using the data at hand to make predictions about

the population from which those data were taken. Jacobs rightly notes that "statistical inference is almost by definition imperfect – all sampling introduces some error" (1997, p. 2). The context of media polls has been used to research how students expect, notice, and understand the variation inherent in sampling (Watson, Collis, & Moritz, 1995; Watson, 1997; Watson, 1998; Watson & Moritz, 2000a; Watson & Moritz, 2000b). Polls invite people to consider the effect of sample size on variation, as well as the variation naturally arising from polling different sample of the same size. One task, involving polls on handgun use, generalized from a sample of 2508 Chicago high school students to make a claim about all high school students in America. Another poll involved listeners who phoned a youth radio station to voice opinions on drug use. The research tasks asked whether or not the samples for these polls offered a reliable way of finding out public opinion throughout the country. Watson and Moritz (2000b) hypothesized a model of student development of concepts of sampling which suggests that

> As students begin to acknowledge variation in the population, they recognize the importance of sample selection, at first attempting to ensure representation by predetermined selection but subsequently by realizing that adequate sample size coupled with random or stratified selection is a valid method to obtain samples representing the whole population (p. 63).

The concepts of variation and representation are intrinsic to the task of making inferences about populations from sample data, and the tension between these two concepts "always exists in a sampling situation" (Shaughnessy et. al., 1999, p. 7).

Rubin, Bruce, and Tenney (1991) agree that a key to mastering statistical inference is to balance sample representativeness (the way in which a sample often has

characteristics that are identical to the parent population) with sample variability (the idea that different samples from a single population are often not identical). To investigate these concepts, Rubin et. al. (1991) interviewed a dozen high school seniors who had never taken any statistics courses. The researchers used a question in which the population was known, and repeated samples could be drawn. In the Gummy Bears problems, students were told that packets of candy were filled with 6 Gummy Bears per packet. These candies were packaged after being drawn from a large vat containing two million green and one million red candies. Students were first asked about the number of green candies they thought would be in their own packet; then they estimated how many packets out of 100 would have that same number of green candies. "Finally, we asked them to specify the entire distribution by answering the questions, 'How many kids out of 100 had N green gummy bears in their packet?' for N = 0 through 6" (Rubin et. al., 1991, p. 5). This is an excellent question to get at variability in repeated samples of a fixed size. For the first question, all twelve subjects said that they would expect 4 out of the 6 candies in their own packet to be green, and their explanations indicated that they were using a ratio approach to this question. However, "when asked if every kid's packet would contain 4 green Gummy Bears, all of the students knew that there would be variation among samples" (p. 5). Some students felt a need to determine a cause for this variation, suggesting that the candies might have gotten stuck together; others clearly felt that any number other than four, while possible, was an example of a flaw in the sample. In looking at the distribution of 100 packets, students consistently overestimated the frequencies near the middle of the distribution, and underestimated the frequencies

near the tails. "No student's distribution contained a peak at a point other than 4G, 2R, and…only two students allowed the possibility of a category being empty" (p. 7), the researchers noticed. For this task, students seemed overwhelmingly influenced by the notion of sample representativeness.

Rubin et. al. (1991) took the same subjects and asked them another question that involved the underlying binomial structure in the Gummy Bears problem. The subjects were asked to imagine that there were only enough lockers at a school for half the children. In order to determine who would or would not receive a locker, the school principal put slips of paper into a bowl. Half of the slips permitted the holder to a locker, and half denied the holder a locker. On the first day of the drawing, three friends pulled three slips and they all got lockers, but on the next day three more friends pulled three slips and all were denied a locker. Rubin et. al.'s subjects were then asked what kind of evidence should be gathered to determine whether or not the slips were properly mixed in the bowl.

Students' responses for estimates of needed sample size ranged from 50 to 500. For example, one student suggested that if the evidence showed 90 Yes Lockers and 5 No Lockers the first day, combined with 5 Yes Lockers and 90 No Lockers the second day, that would be good evidence of unfair mixing. The researchers comment that "students consistently chose samples that were extremely unlikely, i.e. likely to occur much less often than 1 out of 1000 times" (Rubin et. al., 1991, p. 9). Students were reasoning in this case as though "sample variability were the most relevant fact about

sampling" (p. 11). These students insisted on a very convincing sample before inferring anything about the population. Thus, in two different contexts, Rubin et. al. were able to powerfully illustrate the twin ends of the continuum between sample representativeness and sample variability. On one end, sample representativeness is the idea that a sample may have characteristics that are identical to the parent population. On the other end, sample variability is the idea that different samples from the same population are not all identical and therefore do not exactly match the population. They conclude by noting that students "lack experience thinking in terms of a *distribution of samples* generated from a particular population" (1991, p. 12, italics added).

Shaughnessy (1997) shares anecdotes about a task in which repeated samples of M&M candies, each of size 20, were drawn from a population known to contain 40% brown candies. He notes that "no one has yet said 'you will get 8 browns every time" (p. 7). His point is that the idea of a range of likely values is accessible to students. Moreover, questions about the likely spread of values in a data set, or about the likelihood of a certain spread reoccurring by repeating the experiment, are good ways to get at the variability inherent in a resampling situation (Shaughnessy et. al., 1999; Shaughnessy, 1997).

The Gumball Task on the 1996 NAEP was a missed opportunity to look at student responses to questions about variation (Shaughnessy et. al., 1999). In the NAEP Gumball Task, students were shown a picture of a gumball machine and informed that the mix of 100 gumballs inside comprised 20 Yellow, 30 Blue, and 50 Red. The question asked students to predict the number of red gumballs that would

occur in a sample of size 10. The percentage of student responses that fell in the top level of the scoring rubric for this question was a quite low 8% (Zawojewski & Shaughnessy, 2000). A troubling aspect of this task is that the question "tries to tap children's understanding of centers but does so in a context which more naturally deals with spreads" (Shaughnessy et. al., 1999, p. 8). Notice for instance, that while the NAEP question is much akin to the first question asked by Rubin et. al. (1991) in the Gummy Bear problem related earlier, Rubin et. al. extended the line of questioning to get at the variability of results of repeated samples.

Researchers have explored ways of expanding the Gumball Task so that it offered respondents a chance to demonstrate what they knew about variation (Torok & Watson, 2000; Reading & Shaughnessy, 2000; Shaughnessy, 1999; Shaughnessy & Ciancetta, 2001). In the amendments to the original 1996 NAEP gumball item (later called the Candy Task for research in America and the Lolly Task in Australia) several different ways of framing the task were created for a study involving 324 subjects from grades 4, 5, 6, 9, and 12 (Shaughnessy et. al., 1999). The situation was changed to a repeated sampling problem in which five samples, or pulls, of size ten were drawn (with replacement). The RANGE version of the question, asked for the lowest and highest number of reds that would result from the five pulls. The CHOICE version provided five preselected lists of possibilities for consideration, and the LIST version allowed the respondents to simply write down the five estimates for the number of reds in each pull. The subjects were randomly assigned to one of the three versions of the task. Results were categorized on the basis of how the students' answers reflected

their sense of center as well as their sense of spread.

A subset of the younger students in the study was given the Candy Task both before and after they did the actual experiment in the classroom. Students took turns making the various pulls, recording the results, and then remixing the contents of the container. The researchers found that there was "considerable improvement in the students' responses after they actually did the experiment" (Shaughnessy et. al., 1999, p. 15), meaning that more students properly incorporated centers and spread into their responses.

Reading and Shaughnessy (2000) extended the Candy Task, altering either the number of pulls or the sample size. For example, they asked students for the numbers of reds if six people each pulled out samples of size 50 with replacement. This allowed for an exploration of responses for an increased sample size. They also asked students to describe the results of 40 pulls, each of sample size 10, and then asked students to graph the results for 40 pulls. Finally, they altered the candy mixture itself from 50R, 20Y, 30B to 70R, 20Y, and 10B. Six elementary and six secondary students were interviewed on these tasks. In all cases the students were asked to provide an explanation for their responses. Results showed that students were better at describing reasons for their responses when talking about centers than when talking about variation. Also the researchers found that "the LIST form of the question appears to give more information about variability than the CHOICE or RANGE versions," and that "it may be hard for students to describe variation with only six handfuls" (Reading & Shaughnessy, 2000, p. 7). Another finding was that the

different mix of colors did not seem to affect student ability in predicting outcomes, although "it appears to be more difficult for students to justify their responses with the 70% mix" (p. 8). The researchers revised their protocol and asked students to imagine pulling 100 samples of size 10, and to draw a histogram for the frequency of reds in each sample. Reading and Shaughnessy suggest that a computer simulation would be useful to display to students. Researchers could then investigate whether students would want to revise their own suggested histogram after seeing simulation results.

It seems fair to wonder what role graphicacy plays in the Candy Task. Torok and Watson (2000) expressed surprise at the general lack of facility of students in generating graphs to show the outcomes of 40 draws of 10 candies each. Shaughnessy and Ciancetta (2001) asked 31 secondary mathematics students to graph the expected outcomes of 100 draws of 10 candies each. Shaughnessy and Ciancetta note that "in general, constructing a bar graph to represent the results proved a difficult task for these students" (2001, p. 13). Torok and Watson (2000) propose an amendment to the Candy Task protocol which would ask students to fill in a partially completed bar graph, suggesting that "this would be likely to reveal more about students' conceptions about the clustering of results around their expected values" (p. 164).

The Candy Task seems well-suited for investigating not only the effects of sample size, but also the effects of increasing the number of samples of a fixed size. Moreover, the kinds of questions which have been asked require facility in reasoning about centers, spreads, and also graphs. There are also ways in which the subjects' sense of distribution can be tapped. Saldanha and Thompson (2001) had groups of

students draw random samples from populations whose composition was not revealed to the students. After noting that the variability among samples made it difficult to make claims about the population composition, the students wanted to look at collections of samples. Thus, "each group drew 10 samples of equal size from a population of objects and the class investigated how these collections, as a whole, were distributed" (Saldanha & Thompson, 2001, p. 2).

By incorporating so many different aspects of statistical thinking, namely centers, spread, and graph sense, the sampling items presented thus far seem very versatile as a way to investigate students' thinking about variation. Research shows that sampling environments provide opportunities to look at the effect of sample size on variation, as well as the way that samples of the same size differ from one another and provide different pictures of the underlying population. A last context for looking at conceptions of variation is in probability situations, and this context is described in the next section.

Variation in Probability Situations

Much of the previous research on sampling is colored by probabilistic thinking. One can imagine being asked for the probability of getting a certain number of red candies in a sample, or the chance of being selected to participate in a survey. For example, I chose to separate Truran's (1994) research on children's understanding of variance from the body of literature discussed so far in this paper, because his study was framed in terms of "one probabilistic situation" (p. 2). Truran used colored balls in an urn, much akin to the Candy Task, but he only had two greens and one blue in

the urn. His subjects, four girls and four boys in each of grades 4, 6, 8, and 10, drew one ball at a time with replacement. Truran based a series of questions on this format, "If we did this again m times would you be surprised if you got n greens / blues?" (p. 3). This is similar to the Candy Task, except that here the sample size is one. This distinction makes the probability aspects of this task more transparent than the sampling aspects. Still, in asking first about 9 draws and then about 50 draws, Truran's aim was to find out about students' conceptions of variation, and in particular to see what range of results the students would consider normal. It is interesting to note that while the protocol asked for students to provide a specific number they would find surprising, in fact some students explicitly talked about *ranges* of surprising values without being prompted. Truran notes that almost all of the subjects "had some awareness that extreme numbers would be surprising," and seven of them "distinguished 'surprising' from 'very surprising'" (1994, p. 7). He also noted some reliance on the availability heuristic, and claimed that the students' naïve understanding of variance depended on their number sense and their facility in computation.

Shaughnessy (1997) mentions that when students are given a probability question that involves the likelihood of a single event, some students may try to superimpose the idea of a sample on the problem when none exists. For instance, a question was posed to middle school teachers in which a fair coin was flipped five times. Teachers were asked what is more likely to occur, A) HTTHT, or B) HHHHH. Some of the teachers responded that "'In a small sample anything is possible' and "the

long term results gravitate towards A),' as if there were an ongoing sample" (Shaughnessy, 1997, p. 3). In fact, there is no sample at all in this question. There is a sample space in the probabilistic sense, but no actual sample from a population is being drawn (Shaughnessy et. al., 1999). Questions like these lead subjects to "focus on single outcomes, rather than a range of possible outcomes" (Shaughnessy, 1997, p. 3). Konold (1995) used simulations to look at many trials of five flips, where the number of trials corresponded with what he called the sample size. He found that students could see the kind of variability among repeated samples of the same size. Thus, the outcomes in one distribution of 1000 samples (with each sample comprised of five flips) will vary from those in another distribution 1000 samples. Konold claimed that "the different outcomes of each repetition reveal the variability inherent in the sampling process and give some sense of the magnitude of that variability for the given sample size" (1995, p. 209), referring to the variability within and across a distribution.

Shaughnessy and Ciancetta (2001, 2002) used a pair of spinners with 31 secondary mathematics students. Each spinner was half black and half white as in a 1996 NAEP task. First the students were asked a pure probability question: "A player wins a prize only when *both* arrows land on black after each spinner has been spun once. Jeff thinks he has a 50-50 chance of winning. Do you agree?" (Shaughnessy & Ciancetta, 2002, p. 2). Then, students were asked to predict the number of times out of ten spins that both spinners would result in black. After predictions were made, students gathered data in sets of ten spins each. After each successive set of ten spins,

students were given an opportunity to revise their predictions. This same protocol was also repeated with a pair of spinners in which the first spinner again was half black and half white, while the second spinner was one quarter black and three quarters white. For some students, there was conflict "in trying to resolve what their 'theories' would predict, and the variability in their sample data" (Shaughnessy & Ciancetta, 2001, p. 12). Prior to gathering data, very few subjects actually listed a sample space for these problems. However, actually performing the experiment, gathering the data, and "seeing the variation in repeated samples of ten trials, led a number of our students to construct the sample space for the spinner problem" (Shaughnessy & Ciancetta, 2002, p. 5). These interview results support the connection between the two concepts of the sample space in probability and the expected variation in values of a random variable. They note that

> The conceptual root of the pedagogical power that we gain from having students conduct simulations is the connection that they can make between the observed variation in data in repeated trials of an experiment, and the outcomes that they expect based on a knowledge of the underlying sample space or probability distribution (p. 6).

Thus, probability experiments offer promise as a viable context for gathering data on people's conceptions of variability. Repeated trials of a probability experiment can focus attention on the variation inherent in the outcomes, rather than just on the expected value for any particular outcome.

Initial Conceptual Framework

Just as variation is at the heart of a statistical investigation, so too is the

understanding of variation at the heart of this study. My primary research question concerns the components of a conceptual framework that can help characterize EPSTs' thinking about variation. At the outset of the study, before designing the tasks and class interventions discussed in the next chapter, I developed a rudimentary initial conceptual framework that I'll describe in this section . Subsequent chapters will show how I used the data from this study to revise and extend this initial framework into what I call an "evolving framework", a richer structure with added depth that addresses my primary research question more fully.

The initial conceptual framework is a synthesis of the ideas promoted by other studies that have looked at variation along with my own ideas based on past experiences studying and teaching stochastics. There are three different "aspects" of understanding variation in my initial framework (see Figure 6): *expecting*, *displaying*, and *interpreting* variation.

Initial Conceptual Framework

[1] Expecting Variation
- A] Describing What is Expected
- B] Describing Why (Reasons for Expectations)

[2] Displaying Variation
- A] Producing Graphs
- B] Comparing Graphs
- C] Making Conclusions about Graphs

[3] Interpreting Variation
- A] Defining Variation
- B] Causes of Variation
- C] Effects of Variation
- D] Influencing Expectations and Variation

Figure 6 – Initial Conceptual Framework

Each of the three aspects has what I am calling different "dimensions" that help define the respective aspect. For example, *expecting* variation has two dimensions: describing *what* is expected and describing *why* it is expected.

Taken together, the aspects and dimensions are a working hypothesis for a framework that might broadly characterize student responses. I conducted a pilot study to test the validity of this framework, and I'll use some examples from that pilot study to briefly describe my original thinking about the aspects and dimensions in the framework.

Expecting Variation

I had originally thought of the expectation of variation as a way for students to demonstrate their intuitive, experientially or mathematically based reasoning in situations for which variation is inherent. Two dimensions that are a part of this aspect include describing *what* is expected and describing *why* it might be expected.

Describing What: To illustrate, when drawing repeated samples of the same size from a population it is useful to think ahead of time about *what* variation is reasonable to expect. Particularly illuminating are the ends of the spectrum. On one end, a person may expect their samples to look very much alike due to a small amount of variation from sample to sample. On the other end, a person may expect huge disparities between samples owing to large amounts of expected variation. In the pilot for this study, a class of elementary preservice teachers investigated sampling variation using bags of M&Ms. I brought in enough bags of M&Ms so that every pair of students could share one bag. The bags were the 1.69-ounce size which are usually

sold individually in stores, and which tend to have between 50 and 60 pieces of candy in a mix of colors. The question posed to the students was, "What's inside a typical bag of M&Ms, both in terms of total candies and in terms of color distribution?" Some students, who had no knowledge of the nature of the way these candies are packaged, anticipated no variation. They felt the bags should not only have the same number of candies, but the same color distribution. Other students expected that the totals and color distributions in each bag would be different, but were unsure of the amount of variation. The work of Rubin, Tenney, and Bruce (1991) supported my consideration of the M&M task as useful in addressing *what* variation is expected.

Describing Why: A separate but related line of questioning concerns reasons for expectations. Thus, a second dimension in the *expecting* aspect is describing *why*. For example, consider a chance situation whereby the data is a compilation of the outcomes of repeated events. The expectation of variation in such a situation can be tied to the research on probabilistic thinking. Reliance on the outcome approach or on proportional reasoning could result in minimal attention to variation because subjects might focus on a single number to represent their best guess. A spinner which is three-fourths black and only one-fourth white might be expected to produce mostly black outcomes, but what variation would students expect, and why? That is, in 20 spins would students expect all the outcomes to be black, and why or why not? Shaughnessy and Ciancetta (2001) considered similar questions with school students.

Expecting variation to occur is fundamental to an overall understanding of the concept, and it is one aspect that was researched in my study. Before conducting any

experiments or attempting a task, and prior to looking at data relevant to the situation, it should be asked, "What variation is expected?", and also it should be asked: "Why?"

Displaying Variation

The three dimensions within this aspect all relate to graphs: Producing graphs, evaluating and comparing graphs, and making conclusions about graphs.

Producing Graphs: Concerning the first of the three dimensions, although I did not find many references in the literature, it seemed to me that producing (making) graphs affords students another way of showing what kind of variation they either see or expect to see in the data. My interest was in the variation the students did or did not reveal in their graphs, and yet I suspected that the students' graph sense would be a major factor in this dimension. That is, there are many types of graphs, and they reveal or obscure variation within the data to different degrees. The kinds of graphs students produce depends on how fluent students are with different graphs types to begin with, but I still was curious about what the student-generated graphs might suggest as far as an understanding of variation. In another example taken from a pilot for this study, students were asked to gather data on the amount of money in coins that we had each brought to class. Only two people out of thirty-two total had exactly the same amount of money, while the range went from zero to over ten dollars in change. In graphing how much change students had brought to class, I asked groups of students decide for themselves how to graph the data. Most groups chose histograms,

but the groups used different interval widths. Figure 7 shows a computer version of what we had in class.

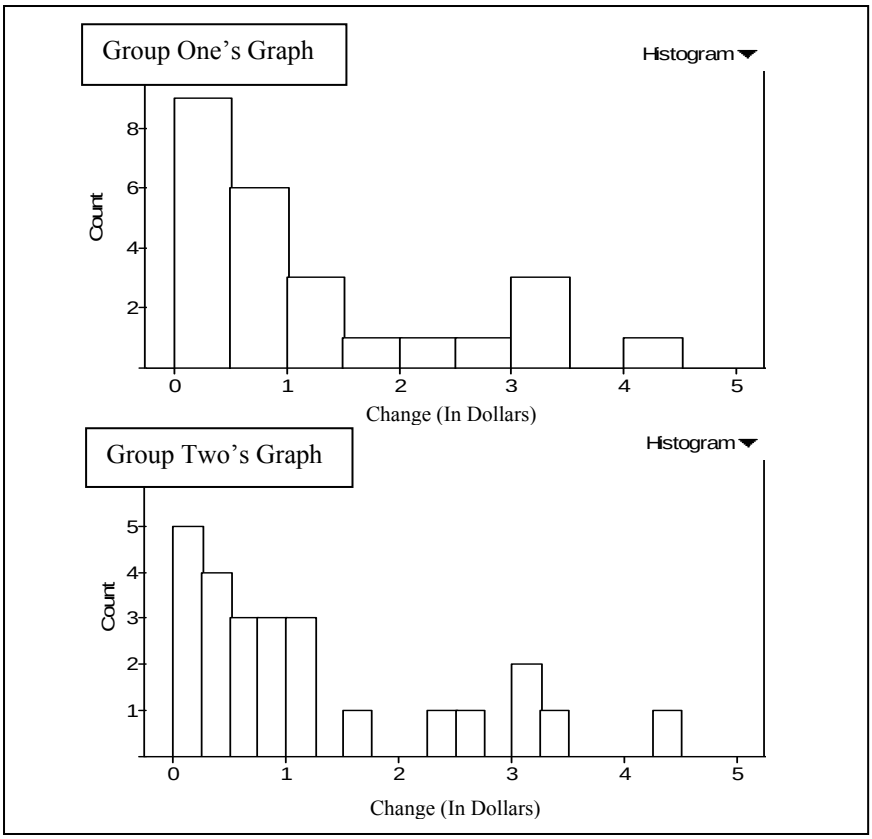


Figure 7 – Change Brought to Class

Interval width plays a large part in determining the extent to which histograms can obscure variation within the data. Some groups of students used 25-cent intervals and other used 50-cent intervals. The graphs told different stories about the variation present in the data, and some students commented on these differences. For example, the larger interval widths disguise the fact that no students had between $1.75 and $2.25 in change.

Comparing Graphs: Whether the students make their own graphs or consider other graphs that are shown to them, I was curious about how they evaluated and

compared graphs with respect to variation (the second dimension within the *displaying* aspect). The M&M investigation from the pilot study offered a good example of this second dimension. Students each prepared a bar chart showing the color distribution of their bag (as a percentage of the total candies in the bag), and all the bar charts were lined up on the chalkboard. The distributions were quite varied. For example, some bags had 20 % red candies and some had 35 % red. Some students didn't seem to notice the variation among bags, and instead seemed intent on trying to find out which color was most dominant across all bags. Other students clearly indicated they could see that the percentages were jumping up and down all across the chalkboard, and that the bags were quite different from one another. The point was that students compared graphs using different strategies, and variation factored into their explanation in different ways.

In another example, as students looked at the results from the data on the amount of change (Figure 7), I was curious to see what comments would emerge regarding the spread of the data. I taught two sections of the course, and later in the week I was able to bring together the results of both sections for comparison purposes. Some students noticed not only that the means of the two data sets were different, but also that the spreads of the data from the means were also markedly different.

Since some graphical displays obscure variation within data more than others, I also wanted to know if students could discriminate between graphs which highlight and graphs which minimize variation. For instance, boxplots do a good job of

showing the range and the interquartile range, but do not show variation of the data within the quartiles. In the pilot study, the class looked at a boxplot representing the percentages obtained for red candies in each bag taken from the M&M investigation, and we had also graphed the same data in a dot plot (see Figure 8).

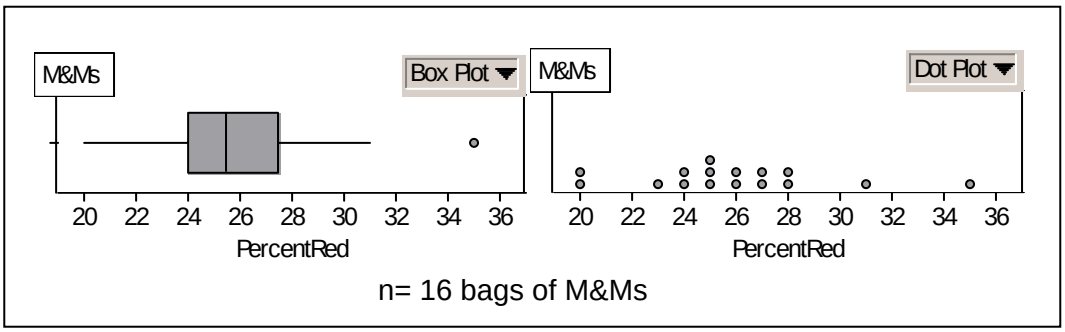


Figure 8 – Percentage of Red M&Ms

The graphs in Figure 8 represent actual data, and made for a good discussion about what gets obscured or emphasized in different types of presentation. The boxplot readily shows the median (25.5) and identifies the middle 50% of the data as falling between 24 and 27.5, neither of which is as quickly obtainable from the dot plot. In the dot plot some counting and calculation must be done to figure out where the quartiles are. Also, the mode is apparent in the dot plot, and the mean can be calculated from information in the graph, but both mode and mean are not obtainable from the boxplot. The range is quickly seen in both graphs, however, the actual distribution of the data – the way values are spread out and vary from one another – is obscured in the boxplot but not in the dot plot. One cannot tell from the boxplot

where the gaps are (such as no values of 21, 22, 29, etcetera), nor can one tell the frequency for each value.

Making Conclusions: I thought of the third dimension in this aspect, making conclusions about graphs, as a natural extension of EPST's reasoning as they evaluated and compared graphs. One type of conclusion I envisioned had to do with questions such as "Which graph shows more variation?" Another way in which I had thought they might make conclusions had to do with which types of graphs were more useful in different situations. As we've seen, research on graphicacy and distributions suggests different graphs provide different degrees of information about centers and spreads.

Interpreting Variation

There are four dimensions within this aspect: Defining variation, causes of variation, effects of variation, and influencing expectations and variation.

Defining Variation: At the core of defining variation is simply trying to find out what variation means to the students. I wondered how they would describe variation, and originally I was expecting both quantitative and qualitative descriptions. Quantitative descriptions of variation that I had seen students use included measures of range, interquartile range, and standard deviation. For example, in the pilot study when comparing two contrived data sets with 25 test scores in each set (see Figure 9), some students commented that the sets had identical ranges. The two data sets also had identical means, medians and interquartile ranges. Although the standard deviations (which are different) were provided to students, not many students commented on this statistic.

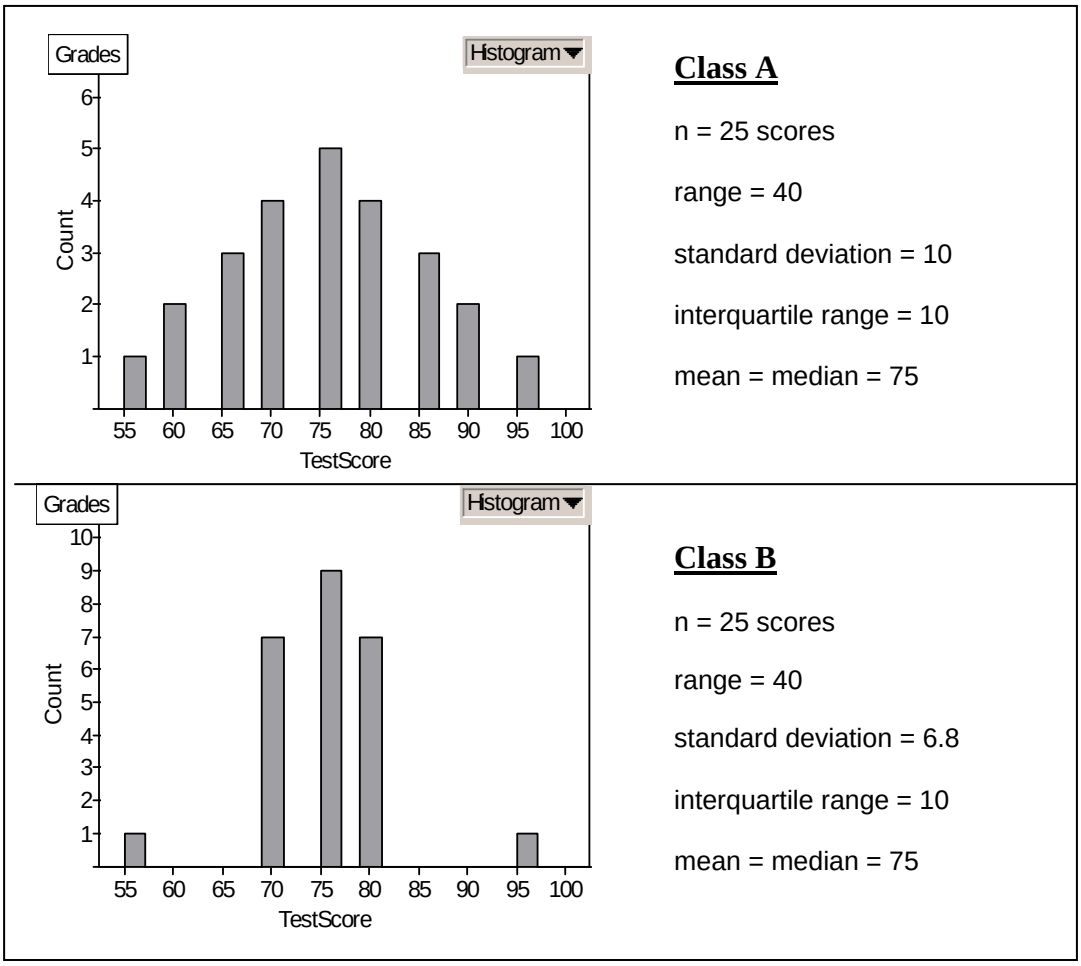


Figure 9 – Class A & Class B

Based on others' research, I believe that EPSTs can have meaningful conceptions of variation independent of any understanding of the standard deviation. Torok and Watson (2000) conducted a study of conceptions of variation with sixteen students (four each from grades 4, 6, 8, and 10). The students responded to questions about sampling and weather data in terms of what they thought was reasonable. Torok and Watson comment that "this study successfully explored students' understanding of variation without ever employing the phrase 'standard deviation'" (200, p. 166). Qualitative descriptions attend to the language that students use as they talk or write

about the data. For example, students may say that the data for Class A in Figure 9 is spread more evenly, or that the data for Class B is clumped together near the center.

Causes of Variation: By asking students about the causes of variation (the second dimension for this aspect), students conjecture and reason about the source of the variation. For example, in the M&M investigation during the pilot study, the students discussed not only why the bags were so different from one another, but also why so few of the bags were representative of the true color proportions that the company claimed to produce. Is there variation among the samples because the company only cares about the weight of the package and not the color mix of each bag? Is there more variation when looking at "fun-size" bags (n= about 24) than "regular" bags (n= about 55) because of the smaller sample size? When conducting a statistical investigation, and in looking at data, it seems useful to wonder about where the variation is coming from and why it is present (Wild & Pfannkuch, 1999). Another interesting line of questioning about causes of variation came from a probability experiment in the pilot to this study. Students repeatedly threw two dice and computed the sums of the numbers facing upward. Theoretically, the probability of obtaining a sum of seven is 1/6. Different groups of students threw dice many times (between 30 and 60), but for each group the percentage of their total throws which yielded a sum of seven strayed either above or below the expected 16.7 %. Students offered reasons for why the experimental percentages did not match the theoretically expected percentage, which included the idea that the dice were uncontrollable, or that luck was not on their side. There was also some class awareness that the variation in this experiment was due to randomness.

Effects of Variation: Thinking about the causes of variation can lead to questions about the third dimension, effects of variation. After doing probability experiments with my class in the pilot study, discussion turned to what the class would predict if some new student joined the class and threw the dice 30 times. Although the probability of obtaining a sum of seven is 1/6, many students argued that 30 throws would not yield exactly 5 sums of seven. Questions of confidence, such as asking "How confident are you that 30 throws would have at least 3 sums of seven?", are questions that probe students' thinking about the effects of variation. If there is already known to be a considerable amount of variation in the situation, as there can be when looking at "fun-size" bags of M&Ms, then extreme results (such as having no Red candies in the bag) seem less surprising. The presence of variation has an effect on inference and confidence, and I was curious how students perceived such effects.

Influences on Expectation and Variation: Finally, the dimension of influencing expectation and variation is something that I first thought of as I considered the Hospital Problem mentioned earlier. In the Hospital Problem, relative size is an issue in the predicted variation. I wondered if EPSTs would think in terms of relative sizes as influencing their expectations of results, or the variation in a set of results, coming from a sampling situation. Also, I wondered if the number of trials performed might be seen by EPSTs as being connected to variation in sampling and probability situations.

Contexts for Understanding Variation

This initial conceptual framework was situated within three contexts for my study of variation: variation in sampling, variation in data sets, and variation in chance

situations. These three contexts arose out of the literature review, and some examples from the pilot study will further clarify the distinctions among these contexts.

Variation in Sampling: The M&M investigation is a good example of how samples vary in the way they reflect the population from which they were drawn. Opening many (smaller) "fun-size" bags shows the effect of taking repeated samples of approximately the same size. Even though the population of M&M candies is reported to be 20% red, a sample of size ten may not have exactly two red. The proportion of reds in the sample may vary widely, but increasing the sample size reduces this variation. For example, ten samples of size twenty have red percentages that vary more than ten samples of size two thousand.

For students who have not done the M&M investigation in a previous class, the population color proportions are unknown. Thus, the sample data is the only information through which an inference about the population can be made. By aggregating data from the individual bags, the effect of an increased sample size on variation between samples can be seen. For instance, four samples of 100 M&Ms usually look more similar to each other than sixteen samples of 25 M&Ms.

Variation in Data Sets: In this context, data could be gathered or provided without explicit ties to either sampling or probability situations. The focus was more on where the data came from, what it meant, and how best to describe the data. For example, when the class investigated how much money in coins we each had brought, the initial point of the exercise was to gather and display data. We did similar exercises by gathering data on body measurements, such as arm span and wrist

circumferences. Once the data was available, then we displayed it and discussed notions of center and spread.

Later in class, instead of gathering data, I provided data sets which were either real or contrived, with the purpose of exploring students' interpretation of the data. Data similar to that shown in Figure 4 and Figure 9 was used to talk about whether one class did better on a test than the other class. In addition to comparing different data sets, I explored different displays for the same data set, using displays which mask or highlight the variability in the data.

Variation in Chance Situations: The probability experiment involving the sum of two dice is an example of this context. The data gathered is completely governed by chance. With the dice experiment, many students knew that the chance of obtaining a sum of seven is 1/6. What is interesting in this context is to provide opportunities for students to move away from answers that rely on an expected value for one outcome, and to move towards the anticipated results of many outcomes in which random variation is sure to play a part. How extreme would the data from tossing the dice have to be before students questioned the fairness of the dice or the legitimacy of the data?

Another in-class probability experiment was done for which none of the students knew the theoretical probability. The task was to write down how many spins of a five-spinner (a disk partitioned into five sectors of equal area) it took until the pointer landed on each of the five numbered sectors at least once. Each student repeated the experiment fifteen times. Since they did not know the expected value for

the number of spins, students could only use their experimental data in making a prediction. The outcomes of only fifteen experimental trials tend to be quite varied.

Summary

Statistical thinking about variability is influenced by a host of different factors. Some of the research deemed most salient to a study of conceptions of variation has been discussed in this chapter, and includes research on randomness, intuitive stochastical thinking, graphicacy, and averages. A lack of appreciation for randomness appears to translate into an inadequate view of variation, and to diminish the promise of statistical inference. Naïve and intuitive views of uncertainty can result in misjudging probabilistic and stochastical situations, including those situations where reasoning about variation is crucial. Graphs are a powerful tool for reducing data, and different graphs convey different degrees of information about center and spread. Also important for data handling are measures of central tendency. All of these ideas relate to conceptions of variation. For example, if students have difficulty reading graphs, then they will likely be unable to notice variation as it is revealed graphically. If they are unable to find a representative measure for a data set, then they will be unable to talk about variation as a spread around the center of a distribution.

In recent years it has become clear that studies crafted specifically to elicit responses on variation can be embedded in contexts which include (although certainly not limited to) sampling, data sets, and probability situations. Although some research has been conducted on students' conceptions of variation in these three contexts, one of the larger gaps in the literature concerns research specifically looking at what

teachers know about variation. Shaughnessy (2002) writes that "we are not aware of any research studies that have dealt specifically with teachers' conceptions of variability" (p. 2), and this paucity of research also holds true for preservice teachers.

Motivated by previous studies and my own experience in teaching probability and statistics, I developed an initial conceptual framework consisting of three aspects of understanding variation (*expecting, displaying,* and *interpreting* variation), along with their corresponding dimensions. The entire framework was situated within the contexts of sampling, data sets, and probability situations. Together, the initial conceptual framework and the contexts for exploring variation provided an organizational structure for the methodology and a coarse lens for the initial look at the data. In the next chapter, I'll describe the methodology for the research and also the process by which the initial conceptual framework became refined.

CHAPTER THREE

Methodology

This chapter discusses the procedures for gathering and analyzing the data. The first section tells *who* the data was collected from, providing a more general description of the research design as well a rationale for *why* methodological components were chosen. The second section offers more detail about the data gathering, including *what* data was collected and *how*. The third section illustrates the specific way that data was analyzed to create a richer framework for understanding EPSTs' conceptions of variation.

Research Design

Subjects were chosen from a section of a course I'll refer to as *Mathematics for Elementary Teachers 2* (MET 2). The course was the second of a two-course sequence at a University located in a metropolitan area of the Pacific Northwest. Intended for prospective teachers, the two math courses (MET 1 and 2) are required for those wanting to enroll in the Graduate Teacher Education Program (GTEP). Completion of GTEP leads to issuance of an Initial Teaching License. To give a sense of the environment common to the subjects, a general overview of the content and pedagogy for the typical MET 2 course is next described. Within the context of the typical MET 2 course, specific features of the particular section used for my research are articulated.

Content and Pedagogy of MET 2

The content for MET 2 includes geometry, probability, and statistics. Instructors usually divide the ten-week course into two parts of roughly equal length, with one part dealing with geometry and the other part dealing with stochastics. The probability component includes single-stage and multistage experiments, and students investigate both theoretical and experimental probabilities. Theoretical probabilities are obtained by deriving the sample space and considering the possible outcomes of a specific event, while experimental probabilities are obtained by simulation. Students consider disjoint events as well as independent and dependent events. The statistics component includes descriptive measures of central tendency (mean, median, and mode) and spread (range, interquartile range, and standard deviation). Students also analyze data using a variety of graphs, such as boxplots, line plots, bar charts, histograms, pie graphs, and scatterplots. Themes of sampling, such as random sampling, stratified sampling, and making predictions based on sampling, are also a part of the statistics component.

The pedagogy for MET 2 varies with the instructor, but there are some common pedagogical themes. One theme is that students participate in activities, both as a class and in smaller groups. The MET 2 classroom is arranged so that students sit in groups, which can be as large as six students per group. Once a problem has been posed, or an activity given, students typically will work singly, then together in their groups, and finally share ideas with each other in a class-wide discussion. The sharing of ideas points to a second common theme, which is communication. Students are expected to communicate what they are thinking about and to ask questions of one

another so as to understand each other's reasoning. A third common theme is that the teacher acts more as an inquiring guide than as a lecturer. Ideally, the teacher facilitates discussion by asking questions, encouraging other students to ask questions, and generally guiding the class to consensus where possible.

Multiple sections of MET 2 are typically offered during any given quarter. During the Winter Quarter when my research took place, classes were held two days a week for ten weeks. Each class session was scheduled to last two hours and fifty minutes. The section I chose to use for my research was led by Steve, an experienced instructor for MET 1, 2 and other courses for teachers and preservice teachers of mathematics. Steve's plan for the curriculum reflected the components described earlier, but he modified his previous practice of doing all the geometry in the first part of the course, and all the stochastics in the second part. Instead, he devoted the first four weeks to geometry, and then gave the next four weeks to stochastics, followed by a last installment of geometry. The modified schedule was done to help accommodate my research plan.

One reason I chose Steve's section is because of Steve's skill in modeling the pedagogical themes mentioned earlier. He lectures less and he poses problems, guides activities, and facilitates discussions. A second reason for choosing his section is because Steve and I had worked well together as colleagues for more than four years, and our philosophies of teaching and learning were similar. The familiarity with each other's styles and congruence of philosophies helped make working together easy, particularly when we were co-teaching some of the lessons. A third reason is because Steve was willing to modify the sequence of geometry and probability and statistics so

that my out-of-class interviews could take place before and after the stochastics portion of the class sessions.

Student Characteristics

The urban setting of the University fosters a wide range of student backgrounds, and it is difficult to describe typical demographic characteristics of the students who take MET 2. Past students in my classes have ranged in age from the early twenties to the middle fifties. Some are undergraduates and some are graduate students. Some want to become teachers and occasionally some are just filling a University math requirement.

The majority of MET 2 students have taken MET 1 at the University, a course taught in a similar style to MET 2. The content of MET 1 includes whole-number arithmetic, number theory, fractions, decimals, and ratios. Other than MET 1, most students have taken few, if any, post-secondary math classes. It may have been as many as twenty or more years since they have had *any* mathematics class. At the outset of MET 1 students often write a "mathography," which entails a description of their past math classes and their feelings about past math experiences. Most students describe themselves as not having been very good at math in the past, and most MET 2 students share negative memories of precollege mathematics. Some students say they feared or hated their math classes, and others say they were bored. The attitude of most beginning MET 1 students is that mathematics is a rule-oriented discipline. The role of the teacher is to reveal the rules, and the role of the students is to memorize and apply the rules.

The MET 1 experience helps to give most students a new and different vision of what "doing mathematics" entails. Because the pedagogy in MET 1 is similar to that in MET 2, most students come to expect a learning environment in MET 2 where they will be active participants and where their thinking strategies are validated. They have been enculturated to the process of problem-solving and communicating their reasoning. Although not all MET 2 students are comfortable with the learning environment, those who have come through MET 1 at least know what to expect.

A total of thirty students completed Steve's section of the MET 2 course. A profile of the background for these students is offered with respect to the following attributes: Class level at the university, gender, when and where the students had taken MET 1, and any prior probability or statistics courses taken. These attributes are summarized in Table 1.

Table 1. *MET 2 Class Profile*

Attribute	Categories of Attribute	Number of Students	Total Students
[1] Class Level at the University	Undergraduate	9	30
	Graduate	21	
[2] Gender of the Student	Male	6	30
	Female	24	
[3] Where MET 1 was Taken	At PSU	26	30
	Not at PSU	4	
[4] When MET 1 was Taken	Within Last Year	24	27
	More Than a Year Ago	3	
[5] Any Prior Probability or Statistics Courses	Yes	12	27
	No or Unsure	15	

The student total in Table 1 is not always thirty because not all data could be gathered for every student. For example, information on the first three attributes was gleaned from the university's information system, and the last two attributes were informed by the PreSurvey which was completed by 27 students during the first week of the course. Appendix B shows the entire PreSurvey instrument.

Almost two-thirds of the class (19 out of 30 students) were continuing on with the same instructor, Steve, for consecutive quarters. This fact is noteworthy because it helps explain the general disposition of the class with respect to their attitudes about studying mathematics. Of the twelve students who could recall having had prior formal education in probability or statistics, eight expressed favourable attitudes on the PreSurvey when asked "How did you feel about Probability and Statistics at that time?" For example, one student wrote that it had been an "interesting class, while another commented that it had been "fairly easy to understand." The four other students with prior formal exposure expressed unfavourable attitudes, such as: "It was my least favorite class in all of college."

When asked how they felt about learning probability and statistics now, only 5 of the 27 respondents expressed explicitly negative thoughts, such as LT, who wrote: "I feel very scared. It makes me nervous." Sixteen students put something that was explicitly positive, such as

DM: "Very excited, looking forward to it."
CS: "Pretty comfortable, ready and excited"
EM: "I'm interested to learn more"

The remaining 6 students had responses that were somewhat neutral in character, as

the comments by GP illustrate: "I'm open to it, but not really excited." It seems reasonable to assume that, since 19 students did have Steve during the previous quarter, there would be some influence on the students' attitudes and expectations. As one such student, MM, put it, "I'm excited about this class because I enjoyed [MET 1]."

Overview of Research Design

The two main sources of data for my study were written instruments given to everyone in Steve's class and individual interviews conducted with six students selected from the class as my case studies. A third source of information was class observations and videotapes made during the stochastics portion of the MET 2 course to help record the learning environment that the students experienced.

Table 2 summarizes the overall research design, the type of information gathered, who it was gathered from, and when it was gathered during the 10-week quarter. The table also shows the contexts (sampling, probability, or data and graphs) emphasized by each instrument or activity.

Prior to doing any class activities in stochastics, during the first week of the academic quarter, baseline information was collected from the MET 2 students in Steve's section. The information was collected via a written survey that addressed prior mathematical experience and contained a range of questions about probability and statistics. Over the next couple of weeks, while Steve taught the geometry portion of the MET 2 course, I interviewed six students outside of regular class time. The interviews were videotaped and lasted about 45 minutes on average. The initial survey

and interview were named PreSurvey and PreInterview because they occurred prior to the stochastics portion of the MET 2 course.

Table 2. *Overall Research Design*

When	What	Contexts	Type of Data	Who
1st week	PreSurvey	All	Written (In class)	Classwide (n=27)
2nd & 3rd week	PreInterview	All	Videotaped (Out of class)	Six cases (n=6)
5th & 6th week	Class Intervention #1 (Four Questions & Body Measurements)	Data & Graphs & Sampling	Videotaped (In class)	Classwide (Varied)
6th week	PostSurvey #1	Data & Graphs	Written (Out of class)	Classwide (n=28)
7th week	Class Intervention #2 (Known & Unknown Mixtures)	Sampling & Graphs	Videotaped (In class)	Classwide (Varied)
7th week	PostSurvey #2	Sampling & Graphs	Written (Out of class)	Classwide (n=30)
6th & 8th week	Class Intervention #3 (Cereal Boxes & River Crossing Game)	Probability & Graphs	Videotaped (In class)	Classwide (Varied)
8th week	PostSurvey #3	Probability & Graphs	Written (Out of class)	Classwide (n=29)
9th & 10th week	PostInterview	All	Videotaped (Out of class)	Six cases (n=6)

In the 5th week of the course, Steve made the transition from geometry to statistics and probability, and I began attending each class session. In addition to making observations, I also videotaped portions of class activities and discussions. Of the many class activities that took place over weeks 5 through 8, six activities were designed as interventions about variation. There were two activities in each of three interventions, with one intervention focused on each of the contexts, data and graphs,

sampling, and probability. The three class interventions are listed in Table 2, and are further described later in this chapter.

After each one of the three interventions had been conducted in class, a take-home assignment was given, which I have called PostSurveys because they occurred after the entire intervention for that context had been completed. The PostSurveys are differentiated according to their context: For instance, the first PostSurvey corresponded to the context of data and graphs.

After the last intervention had happened in class, and the PostSurvey (Probability) had been administered, a second interview was conducted with the same six students as those interviewed for the PreInterview. The second interview was called a PostInterview because it took place after the interventions occurred in class. Like the PreInterviews, the PostInterviews were videotaped and lasted about 45 minutes on average. All PostInterviews occurred in the last two weeks of the quarter.

I used all the data collected to help inform my research questions. The written documents, observations, and interviews formed a corpus of data that I analyzed with grounded theory techniques to describe components of a conceptual framework characterizing EPSTs' thinking about variation. I also used the entire corpus of data to consider comparisons of EPST's thinking before and after the instructional interventions and to consider which tasks were most useful in examining conceptions of variation. In presenting my results, I used a case study format to profile the thinking of the six cases who participated in the interviews.

Rationale for Design

This is a qualitative study that combines two traditions of qualitative inquiry: grounded theory and case studies. I'll justify the choice of these two traditions by describing how they helped answer my research questions, and then articulate the three types of data collected: written documents, observations, and interviews.

Grounded Theory: Strauss and Corbin (1994) said that "grounded theory is a *general methodology* for developing theory that is grounded in data systematically gathered and analyzed" (p. 273, italics in original). I used three techniques from grounded theory - open coding, axial coding, and constant comparison – to theorize an "evolving framework" for characterizing EPSTs' thinking about variation. This evolving framework, which addresses my first research question about finding components of a conceptual framework, is described in the next chapter. I'll next offer a brief description of the techniques and terminology of grounded theory.

Grounded theory begins with *open coding*, a process of describing and building categories of similar phenomena, the dimensions of which are defined by their conceptual properties. Adding to the power of open coding is *axial coding,* defined as "the process of relating categories to their subcategories [or properties], termed 'axial' because coding occurs around the axis of the category, linking categories at the level of properties and dimensions" (Strauss & Corbin, 1998, p. 123). Microanalysis is the combined approach of open and axial coding, often using a line-by-line analysis, to "generate initial categories (with their properties and dimensions) and to suggest relationships among categories" (p. 57). Emergent tentative hypotheses suggest links between categories and properties (Patton, 2001; Merriam, 1998). As

data are iteratively compared to the emerging categories, the categories themselves are refined in light of reviewing the data (Strauss & Corbin, 1994). "This process of taking information from data collection and comparing it to emerging categories is called the *constant comparative* method of data analysis" (Creswell, 1998, p. 57, italics in original). Patton (2001) calls this comparative analysis "a central feature of grounded theory development" (p. 490). Later in this chapter, I'll demonstrate how I used the techniques of grounded theory to derive the evolving framework.

Case Studies: Case studies are often associated with ethnographies, grounded theories, or exploratory research. Creswell (1998) defines "a *case study* [as] an exploration of a 'bounded system' or case (or multiple cases) over time through detailed, in-depth data collection involving multiple sources of information rich in context" (p. 61). Stake (1994) agrees that a case study should represent "a specific, unique, bounded system" (p. 237), and the boundaries can be defined by time and place. The MET 2 course, conducted in the same location over ten weeks with the same students and instructor, represented the kind of bounded system needed to conduct a case study. I chose six students from the class to serve as my cases (how I chose them is described in the next section), and I used the evolving framework to compare their conceptions of variation from before to after the instructional interventions. Thus, case studies helped me address my second research question about comparing EPST's conceptions over the duration of the research.

There are two specific reasons why case studies worked well together with a grounded theory approach for the purposes of answering my research questions. One reason is because case studies, like grounded theory, allow theory to be generated via

descriptive data. Case studies encourage the use of descriptive data "to develop conceptual categories or to illustrate, support, or challenge theoretical assumptions held prior to the data gathering" (Merriam, 1998, p. 38). McMillan and Schumacher (1997) wrote that "case studies are appropriate for exploratory and discovery-oriented research" (p. 395). A second reason is because both case studies and grounded theory encourage the use of multiple sources of data (Stake, 1995; Strauss & Corbin, 1998). The use of different data sources is referred to as *triangulation*, which serves to "clarify meaning by identifying different ways the phenomenon is being seen" (Stake, 1994, p. 241). Triangulation strengthened my research by letting me get information in several forms and at different times, so that my findings were "consistent with the data collected" (Merriam, 1998, p. 206). My study was triangulated by three methods of data gathering: written documents to review (via the surveys), classroom observations during the instructional interventions, and individual interviews with my cases.

Written Documents: The use of written documents as a method for collecting data in a case study is well regarded (Stake, 1995; Merriam, 1998; Patton, 2001). Written documents can supplement observational data, and "quite often, documents serve as substitutes for records of activity that the researcher could not observe directly" (Stake, 1995 p. 68). In my study, there were two main types of written documents I collected: pre-activity documents (the PreSurvey) and post-activity documents (the PostSurveys for Sampling, Data & Graphs, and Probability). All of the survey instruments are listed in Appendix B. I also collected in-class work that came out of the three interventions. For example, when we began the class

intervention on sampling, students wrote about what they thought a "sample" was, who they thought used samples, and why they thought taking samples might be useful. Students also wrote down initial predictions for what they thought would result from thirty samples each of size ten taken from the Known Mixture. Small groups of students produced posters for actual results from the Known and Unknown Mixture activities, and I saved or photographed all the posters.

Observations: The purpose of using observations as a data collection method was to record the overall class contexts in which the activities occurred. I wanted not only to capture the contributions of my six cases, but also to hear what the ideas the rest of the class shared within their small groups and in classwide discussions as they engaged in the activities.

Observations are a common data collection technique in case studies (Stake, 1995; Merriam, 1998). Best and Kahn (1998) write that "when observation is used in qualitative research, it usually consists of *detailed notation* of behaviors, events, and the contexts surrounding the events and behaviors" (p. 253, italics in original). Patton (2001) includes the following three dimensions of concern when conducting observations: the role of the observer, the disclosure of observation, and the recording procedures.

In my research, I had roles as both participant and observer. I was a participant by virtue of co-directing some of the MET 2 activities, and an observer by virtue of recording the class activities. Patton claims that "the participant observer employs multiple and overlapping data collection strategies, being fully engaged in experiencing the setting (participation) while at the same time observing and talking

with other participants about whatever is happening" (pp. 265, 266). Regarding disclosure, the participant observation is what Fraenkel and Wallen (2000) called *overt*, because the researcher will be identified and the cases will know they are being observed. For recording procedures, I videotaped the three class interventions with the help of a colleague, Matt. I also videotaped parts of all the other class sessions having to do with probability and statistics, and took notes after each session ended. To "minimize the errors resulting from faulty memory," caution Best and Kahn (1998), "simultaneous recording of observations is recommended" (p. 295). Videorecording of the classroom during the activities showed the context of the learning environment, the overall flow of the class, and the specific contributions of my cases to the class discussion. At multiple times during the three class interventions, the specific tables where my cases sat were videotaped, so that I could capture what they were saying to each other in small-group discussions during the activities.

Notes from my observations were added to the corpus of data that was used to help shape the evolving framework, compare EPSTs' conceptions before and after the interventions, and inform which tasks were useful in examining EPSTs conceptions of variation. Thus, observational data helped supplement my thinking about all three research questions.

Interviews: A semi-structured, task-based interview format was the third method used for gathering data. Interviewing is a common and powerful method of trying to understand how other people think (Fontana & Frey, 1994). Best and Kahn (1995) note that "interviews are used to gather information regarding an individual's experience and knowledge" (p. 255), and Patton (2001) says that the purpose of

interviewing "is to allow us to enter into the other person's perspective. Qualitative interviewing begins with the assumption that the perspective of others is meaningful, knowable, and able to be made explicit" (p. 341). By semi-structured, I mean that the interviews were scripted at the outset, but my protocol allowed for a variety of probes depending on the responses of the responses of the interviewees. By task-based, I mean that the subjects were not interacting merely with me as an interviewer, but "with the task environments" (Goldin, 2000, p. 519).

Goldin (2000) goes on to mention the value in task-based interviews, noting that the tasks can be adjusted in wording and content according to the results of previous research. He adds, "Interview contingencies can be decided explicitly and modified when appropriate. In comparison with paper-and-pencil test-based methods, task-based interviews make it possible to focus research attention more directly on the subjects' processes of addressing mathematical tasks" (p. 520).

I conducted the first round of interviews, the PreInterviews, after the PreSurveys had been collected and reviewed but before any of the class sessions in stochastics had begun. The second round of interviews, the PostInterviews, took place during the last two weeks of class, after the three interventions had taken place. For each round of interviews, there were some tasks given which were identical or to tasks given on the survey instruments. Also, the tasks for the PostInterview reflected themes that had been explored in the class sessions. The interviews were all videotaped so that the subjects' explanations and nonverbal communication was recorded. Some of the cases wrote on their copies of the interview scripts, and their written notes were collected when relevant. Their written notes and observations that I

made during the interviews became a part of the overall data for the study, along with the transcriptions of the interviews.

In summary, this research was designed to gather data from multiple sources. All of the data was used to build my evolving framework using grounded theory techniques. The entire framework is explicated as the first part of my results in Chapter Four. Although I did use all the data in comparing EPSTs' conceptions, I focused on data from six cases to exemplify some comparisons of EPST's conceptions. I present comparisons of the six cases as the second part of my results in the next chapter, further organizing the presentation around the tasks that were most illustrative for the research. In the next section, I'll describe specific details of the data gathering.

Data Gathering

This section is organized chronologically. The PreSurvey was administered before the PreInterviews were conducted. Then, each of the three class interventions was followed by a corresponding PostSurvey as shown earlier in Table 2. Finally, PostInterviews were conducted. All the instruments are found in Appendix B

PreSurvey

The PreSurvey was given during the second class session of the first week of the quarter, and was completed by 27 students. The students had known that a survey was to be administered, because I had visited the first class session and told them. Also during the first class session, I described the research project and their opportunity to be involved, and distributed the informed consent forms found in Appendix A. All students were willing to have their written work included as part of

the collected data, and all eventually gave consent to be videotaped during the class sessions. I also had eleven students volunteer to be interviewed outside of class, so I used the PreSurvey responses to help determine who would make up my final six cases (described further in the next chapter). The average time for completion of the PreSurvey was about 45 minutes. The structure of the nine-page PreSurvey had two parts: The first part was the first page, containing background questions to determine the attributes shared earlier in Table 1, such as what prior experiences in probability and statistics they could recall having. Also on the first page were questions about the meaning of the terms "random" and "variation." The second part of the PreSurvey (comprising the other eight pages) held a total of nine questions, many of which had multiple parts. The specific questions and classwide responses for the PreSurvey are summarized in the next chapter, but the contexts for the questions in second part of the PreSurvey are given in Table 3.

Aside from the background questions, all the rest of the questions on the PreSurvey were either identical to or very similar to questions asked by other researchers of middle and high school students. In particular, a NSF-funded project (Shaughnessy, 2003) used written surveys with 12 different classes of middle and high school students, and many questions on the PreSurvey came from the NSF surveys, which in turn had been motivated by prior research. For example, on the PreSurvey and the NSF high school survey, two questions asked students for a description of what the terms "random" and "variation" meant to them, and those questions were similar to the ones used by other researchers (e.g., Watson, et. al., 2002).

Table 3. *Contexts for PreSurvey (Part Two) Questions*

Question	Brief Description	Contexts
1	Results are predicted for drawing one, several, and six samples of 10 candies from a jar (60 Red & 40 Yellow)	Sampling
2	Ranges are predicted for drawing six and then thirty samples of candies	Sampling
3	Results are predicted for drawing fifty samples of candies	Sampling
4	A graph is made to show what the results of fifty samples might look like	Sampling & Graphs
5	Test scores are shown for two different classes. The question addresses the two classes' relative performance	Data & Graphs
6	Two graphs, showing the student heights at two different school, are compared to see which shows more variability	Data & Graphs
7	Results are predicted for performing one, two, and six trials of 50 flips of a fair coin.	Probability
8	Chances of winning a game involving two 50-50 (Black-White) spinners are addressed.	Probability
9	Chances of winning a game involving a 50-50 and a 25-75 (Black-White) spinner are addressed.	Probability

A question involving the comparison of graphs used in the PreSurvey (Question #5) also was used by Watson and Moritz (1999). Question #1, using samples of candies from a jar (akin to the Candy Task mentioned in Chapter 2), was used on the NSF surveys as well as in other research (Torok & Watson, 2000; Reading & Shaughnessy, 2000; Shaughnessy et. al., 1999; Shaughnessy & Ciancetta, 2001). Thus, for the PreSurvey, I chose to use questions that have been used by other researchers to obtain an overview of middle and secondary students' thinking about variation.

PreInterview

After the class in which the PreSurvey was completed, I read through all the responses to get a general sense of how much the students had written and what they had to say. I paid particular attention to the surveys from the eleven students who had said they were willing to be interviewed. Of these eleven students, I selected six who I

predicted would make good cases because I thought their responses on the PreSurvey were representative of the whole class in the sense that their responses were similar to those of other students in the class. Also, my six cases provided enough written detail to convince me that they would have no problem sharing their thoughts in an interview situation.

Having a total of six cases is largely a pragmatic decision, determined by the resources available in conducting my research. Qualitative inquiry generally involves relatively small samples, and the key in purposeful sampling is to select "information-rich cases whose study will illuminate the questions under study" (Patton, 2002, p. 230). Creswell (1998) discusses the value in purposeful sampling, saying that cases can be chosen to provide different perspectives on the phenomena under study.

Because my research design required pairs of Pre and PostInterviews for the same subjects, I conducted the PreInterview of all eleven of the volunteers over the second and third weeks of the quarter. I interviewed all eleven because although I was planning on six cases, I could not guarantee that those six would still be enrolled or be available at the end of the quarter for the PostInterview, nor could I guarantee that the quality of their information in the interview setting would be as rich as I had expected it to be. It turned out that the six cases I initially had in mind from reading their PreSurveys were in fact very useful for informing my framework and helping me understand their conceptions of variation, and they did complete the postinterviews as well. The other five of the eleven volunteers also completed postinterviews, which I conducted because I had the resources and figured that their contributions gave me additional data with which to work on future research.

The PreInterview subjects met with me outside of the regular classtime. A videocamera was set up to record the interview, and a separate tape recorder also was present because the audio cassettes were useful in transcribing the dialogue. I had a copy of the interview script (see Appendix B) and also gave a copy to the subject, who was encouraged to write on the script as desired. The PreInterview contained 13 multi-part questions. The contexts for the questions are given in Table 4. More details of the PreInterview questions and sample responses are provided in the next chapter.

The first question on the PreInterview was identical to the first question on the PreSurvey (akin to the Candy Task) for two reasons. First, I wanted to see how their verbal responses compared with what they had written, and to see if they had anything more to add to their earlier explanations. Second, I wanted a familiar context to ease them into the interviewing mode. The smoothness of the transitions from question to question were important because experience has shown me that sometimes the scenario described in a question can be confusing to students who had never actually done such activities. For example, telling someone to imagine drawing 6 handfuls of 10 candies each, with replacement, from a jar containing 100 candies in it (40 of which are yellow and 60 of which are red), can seem like an overwhelming amount of different numbers to keep track of for one question.

Table 4. *Contexts for PreInterview Questions*

Question	Brief Description	Contexts
1	Results are predicted for drawing one, several, and six samples of 10 candies from a jar.	Sampling
2	Lists are shown for different outcomes of six trials. Subjects are asked to comment on the likelihood of each list occurring.	Sampling
3	The supposed results of 30 samples are shown. Subjects are asked about the likelihood of the results being real or fake.	Sampling & Graphs
4	The supposed results of 300 samples are shown. Subjects are asked about the likelihood of the results being real or fake.	Sampling & Graphs
5	Three graphs show different ways of portraying the same data set. Subjects are asked how the graphs differ.	Data & Graphs
6	A set of 21 measurements for the duration of a train ride is given. Subjects are asked for reasons why the results are not identical.	Data & Graphs
7	The 21 measurements from Q6 are graphed in two different ways. Subjects are asked to compare the two graphs.	Data & Graphs
8	Two different graphs are shown: Wait-times for eastbound and westbound trains. Subjects are asked to compare the graphs.	Data & Graphs
9	Results are predicted for one sample of sixty tosses of a fair die..	Probability
10	Supposed results are shown for four samples with the die, and subjects are asked which results seem real or fake.	Probability
11	Results from repeated samples with the die are predicted.	Probability
12	A 2:1 (White:Black) spinner is used, and results are discussed in terms of what seems surprising to the subjects.	Probability
13	For the spinner in Q12, one sample is defined as sixty spins. Two graphs showing supposed results from 20 samples are compared	Probability & Graphs

Because the PreInterview took place before the stochastics portion of MET 2, I did not want to assume that the questions would automatically make sense to the subjects. Therefore, just as the PreSurvey was scaffolded to include one sample, then several, then six, so too was the PreInterview modeled to move gradually to ever larger numbers of samples.

The PreInterview script contained specific questions that I used with each subject, but the protocol also allowed me flexibility to follow the subjects' train of thought. Thus, each interview contained common questions as well as unique input from the different cases.

Class Intervention #1

All of the class interventions are described in greater detail in Appendix C, so in this chapter they are only briefly discussed. The two activities comprising the Class Intervention for the context of data and graphs were called "Four Questions" and "Body Measurements".

The "Four Questions" activity was chosen for two reasons. One reason is because Steve and I had each used versions of the activity with other MET 2 classes, and were therefore experienced in how it went and what it offered. The second reason is that it offered a good opportunity to discuss both average and spread in data sets. Steve therefore started the class exploration of statistics in the fifth week by having the entire class gather data from one another in response to four questions:

> How many pets do you have?
> How many years have you lived in Portland (to nearest half-year) ?
> How many people are in your household?
> How much change (in coins) do you have today?

After graphing the data in different ways, the class had a discussion about levels of detail provided by each type of graph and about what were "typical" values for a MET 2 student or for the whole class. The tension between centers and spread of data was one theme to emerge from the discussion over graphs from the "Four Questions" activity.

The second activity in the class intervention that focused on the context of data and graphs was "Body Measurements", which was also selected because it was a well-rehearsed activity for Steve and me. More importantly, a similar activity was to be used for the NSF-sponsored project with middle and high school classes, and

comparisons could be made between EPST's and precollege students in the future. As in "Four Questions", we gathered class data for "Body Measurements": Everyone's own armspan, height, handspan, head circumference, and pulse rate per minute were recorded. Also, all students in class measured Matt's armspan, to gather data from a repeated-measurements experiment. Again, we had a class discussion about the data and graphs for the body measurements, this time focusing more on causes of variation.

Data & Graphs PostSurvey

The Data & Graphs PostSurvey, a take-home assignment, was given at the end of the final class session of week 6, and collected the following week. Students had from Thursday to Tuesday to work on the PostSurvey, and they were encouraged to work together on the assignment. I can't be certain who did or did not work in teams outside of class, but traditionally groups of MET 1 and MET 2 students do spend time before or after class working together. The PostSurveys were graded by Steve only as "done" or "not done", similar to other writing assignments given in class. Table 5 summarizes the questions asked on the Data & Graphs PostSurvey.

In creating the Data & Graphs PostSurvey questions, I wanted to examine students' reasoning as they compared or evaluated different graphs. For example, would students refer more to centers or to spreads? Also, I wanted see what kinds of causes for variation they could come up with on their own. In Q1c, I was curious about their ability to reason from the given average of 4 inches of rain to generate a reasonable graph showing appropriate variation for the rainfall during a typical June in Columbus.

Table 5. *Summary of Data & Graphs PostSurvey Questions*

Question	Brief Description
1a	Bar charts are given showing the 30-year average monthly rainfall for Portland and Columbus. Students discuss differences and causes in the rainfall patterns.
1b	Boxplots are given for the same data sets used in Q1a. Again students discuss differences, and also are asked which city they think is rainier, and why.
1c	Assuming that the average June rainfall in Columbus is 4 inches, students are asked to draw a graph showing what each day's rainfall in June might look like.
2a	Dotplots and boxplots are given showing annual traffic death rates for two regions in America, the South and Northeast. Students are asked to compare the rates.
2d	Students are asked to think of factors that might explain the differences in rates between the two regions.

Finally, for my six cases, and the others who would be interviewed a second time, I wanted additional evidence in their reasoning about variation in histograms, dotplots, and boxplots.

Class Intervention #2

In the seventh week of class, the two activities "Known Mixture" and "Unknown Mixture" were done with Steve's students. Matt and I had done this activity at six schools as a part of the NSF-sponsored project. We had seen how effective the activities could be in drawing attention to variation. For example, middle and high school students always commented on the different ranges of the graphs depicting actual results of the sampling activities.

Prior to the Known Mixture, we started with a general discussion of what samples were, who uses samples, and what samples were good for. Then the scenario in Figure 10 was given as a part of a handout. The class discussed initial expectations for this scenario, especially focusing on what would happen if the random draw of 10 names were to be repeated thirty times.

Scenario for Known Mixture Activity

The band at Johnson Middle School has 100 members, 70 females and 30 males. To plan this year's field trip, the band wants to put together a committee of 10 band members. To be fair, they decide to choose the committee members by putting the names of all the band members in a hat and then they randomly draw out 10 names

Figure 10 – Known Mixture Activity

After students talked about predictions for drawing thirty samples each of size ten, we simulated this activity using chips in a jar. Actual data was gathered and graphed. Then we had a discussion about how the graphs of the predicted data compared to one another, how the graphs of the actual data compared to one another, and also how the predicted graphs compared to the actual graphs.

We then made a transition into the second activity in this intervention, the Unknown Mixture. It was made clear that even though we had known what was in the earlier jars, samples still had varied. Now we had larger jars, each containing 1000 chips of yellow and green with the same mixture. However, the exact mixture was not known to the class (it was actually 550 yellows and 450 greens). The students were asked to decide in their groups what sample size they wanted to use (we imposed an upper limit of size twenty for all groups) and how many samples they wanted to draw. Then they were to carry out their plans, do the sampling, graph the results, and make some conjectures about the true mixture in the jar. After the simulation was carried out, we had a class discussion about the different choices made in sampling, the class results, and we tried to forge a class consensus about what the true mixture was.

Sampling PostSurvey

The Sampling PostSurvey take-home assignment was given at the end of the final class session of week 7, and collected the following week. Table 6 summarizes the questions asked on the PostSurvey.

Table 6. *Summary of Sampling PostSurvey Questions*

Question	Brief Description
1a	A boxplot shows the results of 20 samples of size 10 drawn from the smaller candy jar (60 Red & 40 Yellow). Individual sample results are then inferred.
1b	Since the minimum result on Q1a shows 3 Reds in a handful of 10, the number of trials needed to get a 0 or 1 Red are predicted.
2	Results are predicted for drawing one, several, and six samples of 100 candies from a larger jar (600 Red & 400 Yellow)
3	Ranges are predicted for drawing thirty and then three hundred samples of 100 candies
4	Results are predicted for drawing fifty samples of 100 candies

I asked Q1a because it got at the idea explored in class about how boxplots can obscure variation. I was also curious to investigate their ability to work with boxplots, especially working backwards and predicting what the underlying distribution might be which led to that boxplot (as opposed to taking the data and creating the boxplot as is typically done). Q1b was added to draw on our class experience with the ProbSim software and the numbers of samples needed to get extreme results. Questions 2, 3, and 4 all were parallel to questions previously asked in the PreSurvey and PreInterview for the Small Jar (60 Red & 40 Yellow), only now the sampling was done from the Large Jar (600 Red & 400 Yellow). I was curious to see how answers for the Large Jar compared to what students had put for the Small Jar, and I also

wanted to have my six cases thinking about the Large Jar because I built additional questions on that sampling scenario into the PostInterview.

Class Intervention#3

There were two activities that made up this intervention, "Cereal Boxes" and "The River Crossing Game". These were chosen specifically because of the probability aspects involved in the activities. Cereal Boxes relies on the use of spinners and River Crossing on the use of dice as random generators, and these two activities were the main ones done in MET 2 involving random devices.

Cereal Boxes actually took place in the first class session of week 2, just before we gathered data for Body Measurements. As explained earlier, there was considerable overlap in the three contexts, and Cereal Boxes is a good example of this overlap. Cereal Boxes is sample-until scenario, assuming that any of five different stickers can be obtained within each box of cereal, and that the five stickers have equal chances of being obtained. The question is, about how many boxes would need to be opened to obtain all five stickers. The situation can be simulated by using an equal-area five-region spinner. Cereal Boxes brings together probability, sampling, and data and graphs in a way that highlights variation.

The second activity for this intervention, the River Crossing Game, involved finding the sum of two dice. Both the Cereal Boxes activity and River Crossing Game are part of the *Math and the Mind's Eye* curriculum (Shaughnessy & Arcidiacono, 1993).

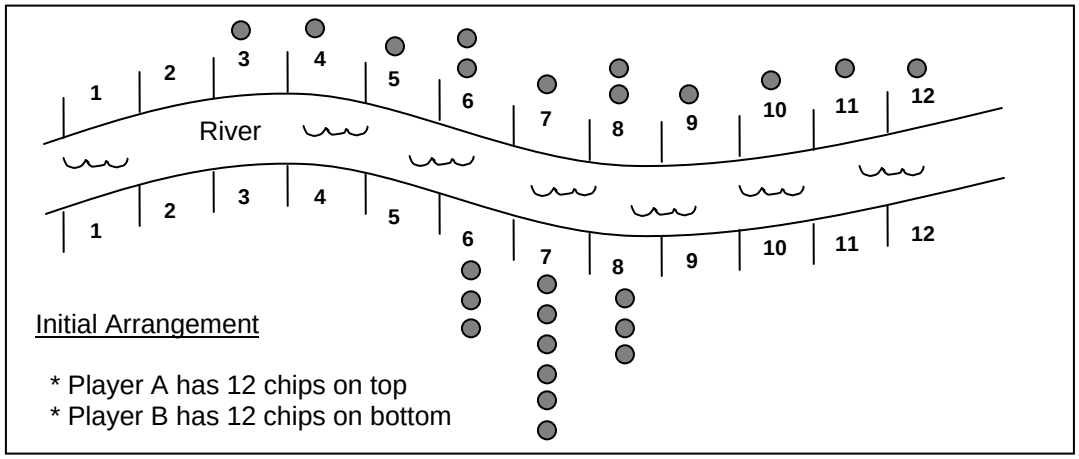


Figure 11 – River Crossing Game

Using two players, each player receives 12 chips to place on their side of a "river", along spaces marked 1 through 12. After configuring their chips in an initial arrangement (see Figure 11 for an example of two players' initial arrangements), players took turns tossing a pair of dice. If either player had any chips on the space showing the total for the dice, one chip could "cross the river" and be removed from the board. The winning player was the first one to remove all the chips on his or her side. For instance, in Figure 11, if the dice resulted in a sum of 10, Player A on top could remove one chip. If the dice showed 8, Player A and B could each remove one chip. As with Cereal Boxes, in the River Crossing Game we made predictions, gathered and graphed data, and discussed results.

Probability PostSurvey

The Probability PostSurvey take-home assignment was given at the end of the final class session of week 8, and collected the following week. Table 7 summarizes the questions asked on the Probability PostSurvey.

Table 7. *Summary of Probability PostSurvey Questions*

Question	Brief Description
1	The number of blacks resulting from 50 spins at a ½ -white & ½ -black spinner make up a single sample. Results are predicted for doing one, two, and six samples of 50 spins.
2	Ranges are predicted for doing thirty and then three hundred samples with the spinner.
3	A graph is made to show what the results of fifty samples (each of 50 spins) of the spinner might look like

Question 1 was crafted to be similar to the sampling questions on the PreSurvey, the PreInterview, and the Sampling PostSurvey, only this time the focus was on probability. Instead of grabbing handfuls of candies, the students were asked to consider samples of fifty spins of a 50:50 (White:Black) spinner. Also, Question 1 was similar to the PreSurvey question on probability that used a sample of 50 flips of a fair coin. Question 2 was similar to what was asked on the PreSurvey and Sampling PostSurvey, and the graph of fifty samples for question 3 was similar to what was asked for fifty samples of the Small Jar in the PreSurvey. I was curious to see how the probability questions with the flips of a coin (in the PreSurvey) compared to the spinner environment (in the PostSurvey), and also how the responses for the probability context compared with those responses for the sampling context. Finally, for my six cases I wanted added familiarity with the spinner scenario, since I based several additional questions in the PostInterview on repeated samples of spinners.

PostInterview

The PostInterviews took place during the 9th and 10th weeks of the quarter, after Steve had gone back to teaching geometry. I followed a very similar protocol with the PostInterviews as I had with the PreInterviews, videotaping each interview as well as making a separate audiorecording for the transcribing process. The PostInterview contained 13 multi-part questions, and the contexts for the questions are given in Table 8. More details of the PostInterview questions and sample responses are provided in the next chapter.

As in the PreInterview, I chose the very first question on the PostInterview to be identical to a question the subjects had already seen (it had been asked in the Sampling PostSurvey. Also, the first four questions were isomorphic to those on the PreInterview, maintaining the population proportion of 60% Red, except that they used the Large Jar (600 Red & 400 Yellow) as opposed to the Small Jar (60 Red & 40 Yellow). Question 5 involved both the Small and the Large jars, the only question in any of the instruments to do so directly. Questions 6, 7, and 9 were asked in part because of their similarity to the MAX train ride questions on the PreInterview. Lastly, question 10 was identical to the first question on the Probability PostSurvey, and set the subjects up for a transition to the last few questions that also involved spinners.

Table 8. *Contexts for PostInterview Questions*

Question	Brief Description	Contexts
1	Results are predicted for drawing one, several, and six samples of 100 candies from a large jar.	Sampling
2	Lists are shown for different outcomes of six samples. Subjects are asked to comment on the likelihood of each list occurring.	Sampling
3	The supposed results of 30 samples are shown. Subjects are asked about the likelihood of the results being real or fake.	Sampling & Graphs
4	The supposed results of 300 samples are shown. Subjects are asked about the likelihood of the results being real or fake.	Sampling & Graphs
5	Two graphs show supposed results of forty samples at the small and also at the large jar. Subjects are asked if graphs are real or fake.	Sampling & Graphs
6	A set of 20 measurements for the weight of a muffin is given. Subjects are asked for reasons why the results are not identical.	Data & Graphs
7	The 20 measurements from Q6 are graphed in two different ways. Subjects are asked to compare the two graphs.	Data & Graphs
8	35 different muffins from the West End bakery are shown. Subjects are asked how much their (36^{th}) muffin might weigh.	Data & Graphs
9	Two different graphs are shown: Muffin weights for East and West End bakeries. Subjects are asked to compare the graphs.	Data & Graphs
10	Results are predicted for doing one, two, and six samples of 50 spins at the 50:50 spinner.	Probability
11	Lists are shown for different outcomes of six samples. Subjects are asked to comment on the likelihood of each list occurring.	Probability
12	A graph shows the supposed results of twenty samples. Arguments from other people about the results are discussed by subjects.	Probability & Graphs
13	Two graphs show supposed results of two classes doing thirty samples at the spinner. Subjects are asked if graphs are real or fake.	Probability & Graphs

It was mentioned earlier how the activities in the interventions were designed to get at variation. As can be seen in Table 8, the activities in the interventions had direct ties to the tasks on the PostInterview questions. For instance, the intervention on data and graphs included different types of graphs and the amounts of variation they showed. Body Measurements got at the ideas behind repeated measurements, as did the muffin weight questions on the PostInterview. The Known and Unknown mixtures had students actually draw chips from a container to experience drawing candies from Large and Small Jars. Cereal Boxes and the River Crossing Game had students use traditional random generators such as spinners and dice to get a sense of

what was likely in a probability context. There is one big difference on the PostInterview compared to the PreInterview: PostInterview questions 8, 9, and 12 all included boxplots as well as either dotplots or histograms, but boxplots were not covered on the PreInterview. Thus, several of the tasks involving graphs had two types of graphs, again relying on the experience gained in the class interventions.

Data Analysis

In this section, I'll illustrate the process of using grounded theory techniques to develop what I call my evolving framework, which addresses my primary research question. The framework in its entirety is presented in the next chapter, which focuses on *results* of the study. The *method* by which the framework was derived is described in this chapter. The process of deriving a framework is laborious and quite detailed, so I'll just use one piece of the framework to provide an example of how I applied the three techniques of grounded theory: open coding, axial coding, and constant comparison. Before launching into my illustrative example, a last methodological component to be discussed is the role of computer software in my data analysis.

Role of Computer Software

The creative yet systematic process of theory building, which in my study took the form of fleshing out a conceptual framework about variation, was aided enormously by the NUD•IST software.

Grounded theory techniques allow for the inclusion of a wide scope of data, such as the written responses, transcribed interviews, and observational notes. As the process of theory development moves through cycles of constant comparison, memos suggesting the continual conjecturing and refinement of categories and concepts also

become a part of the data. Thus, data management becomes a crucial issue in using grounded theory. Richards (1994) boldly states that "all researchers working in the qualitative mode will clearly be helped by some computer software" (p. 105). The use of qualitative data analysis software facilitates not only the management of data, but "it can offer leaps in productivity for those adept at it" (Patton, 2001, p. 447).

The software used in this study was NUD•IST (Non-numerical Unstructured Data Indexing, Searching, and Theorizing), a theory-building program that aids in data storage, coding, retrieval, and category comparisons and linking (Richards and Richards, 1994; Richards, 1994; Patton, 2001). NUD•IST is well-suited for the analysis techniques of grounded theory, although it cannot be emphasized enough that software only assists in the process – software does *not* analyze data for the researcher (Patton, 2001; Creswell, 1998). The flexibility in coding, categorizing, and revising made the development of theory a dynamic and reflexive process. The single biggest obstacle was in learning what NUD•IST could do for the analysis, and how to get the program to do what I wanted. Once the various ways of categorizing, indexing, and coding were learned – the logistics of the program as well as the potential – NUD•IST became an invaluable tool.

Developing the Evolving Framework

As a tradition of qualitative inquiry, grounded theory allows analysis to begin even in the absence of any initial structure or preconceived ideas of what the data might hold. For my research, I did start with an initial conceptual framework, which was based on previous research as well as my own experience. However, the initial conceptual framework was a rough structure that offered little in the way of specifics;

it was mainly used as an overarching guide for designing my tasks. Recall that the initial framework had three aspects: *expecting*, *displaying*, and *interpreting* variation. Each aspect had different dimensions. For example, *expecting* had the two dimensions of *what* was expected and also *why* it was expected. Grounded theory offered a way to expand on the dimensions of the initial conceptual framework. I'll show how this expansion happened with the dimension of *what* was expected.

I'll start with the process of *open coding*, whereby I gathered all responses having to do with *what* was expected and looked for what I called broad common "themes" within the responses. Open coding led to the emergence of three themes for the dimension of *what* was expected: responses concerning the expected value, responses about repeated values, and responses about a range or extreme values. I then used *axial coding* to focus on these three themes in turn to describe them in terms of what I called "characteristics" of the theme. The process of *constant comparison* meant that I iteratively went back and forth from the data to the emerging themes and characteristics, looking for confirming and disconfirming evidence from all my data sources to help me conceptualize the evolving framework as it was being built.

Open Coding: All my observation notes, memos, interviews, and survey data had been transcribed and imported into NUD•IST as text files. I chose to set a line of text as the smallest unit which could be coded (other choices included setting a paragraph or section or the entire document as the unit of analysis). I then went through all the data and coded all the occurrences where I found references to *what* was expected. In NUD•IST the process of coding means highlighting the lines of text

and selecting a "node" to code the text at. Every node has a title, and on this first pass at coding I simply titled my node "Describing What is Expected".

The examples I'll use in this section are responses taken from my six cases from Question 10 on the PostInterview. The process I'll be describing was applied to all of my data, but in order to illustrate the depth of analysis that led to the framework, I need to limit the example. Question 10 had three parts (Q10a, Q10b, and Q10c), and imagined a man getting a sample of 50 spins of a half-black and half-white spinner. The question in Q10a was "How many times do you think the arrow might land on black? Why?" Table 9 shows some sample responses that I coded at "Describing What is Expected" for Q10a.

Each line of text is coded at the level of my dimension ("Describing What is Expected"), but a closer examination of the responses shows different themes within the responses. A review of my memos showed that the first thing I noticed was how both DS and EM used the words "close to 50%", and JM used the parallel "approximately 50%". Comparison of other data for these cases showed me that they did know 50% of the 50 spins in this situation was 25 blacks, so when they referred to "50%" it was isomorphic to saying the expected value. That got me thinking that a theme to look for in their responses had to do with what they said about the expected value.

Table 9. *How many times might the arrow land on black? Why?*	
Subject	Response
DS	Like, just close to 50%, but not exactly! Yeah, within 2 or 3.
EM	I think it will land there somewhere close to 50% of the time, I don't think it will always be 50% of the time, I think it will be probably between 40 and 60% of the time. So, 25...between 20 and 30 spins.
JM	JM Well, approximately 50%, but it will be , you know, plus or minus, maybe, 20% of that number – Somewhere in there.
RL	Oh, how about… Somewhere between 21 and 29… I don't say, you know, 18 to 29… I'm going afar from 25 in either direction. It's probably within that [21 to 29] range.
SP	Yeah, so it expect it , like, up from 25, maybe like 30, and 20 … between 20 and 30

I then coded the responses of DS and EM and JM as "Concerning Expected Value", since that was a theme I was hypothesizing. A strength of NUD•IST is that one can code text at as many nodes as one wishes. Now, for example, DS's response has two codes: one code at the dimensional level "Describing What is Expected" and also at the thematic level "Concerning Expected Value." At this point, I wouldn't know much about the characteristics of the theme "Concerning Expected Value" until I'd done axial coding. Something else I noticed in the responses was a focus on range, such as when DS mentioned "within 2 or 3", or EM said "between 20 and 30 spins". Notice how RL gives both a range he is comfortable with (21 to 29) and a range he thinks is too wide (18 to 29). At this point, I created a node for the theme I was hypothesizing, and titled it "Concerning Range or Extremes". JM's response also reflected the theme concerning range ("...plus or minus, maybe, 20% of that number") as does SP's response ("between 20 and 30"). Lastly, I noticed how EM said "I don't think it will always be 50% of the time," and that made me wonder about a possible

theme concerning repeated values. That is, I thought maybe I should be on the lookout for responses that specifically mentioned if results would be the same or different from one sample to the next. I had a tentative theme "Concerning Repeated Values" which I carried into the next set of responses with my other themes.

This illustrative example omits many of the other memos and conjectures I had in the first pass at open coding, and makes the process appear more streamlined than it was. The point of microanalysis (open coding combined with axial coding and the method of constant comparison) is that you can begin with broad and multiple categories and properties and then winnow them down, collapsing and combining as you continually re-conceptualize your data. Thus, there are many other themes that had occurred to me in my first pass at looking at the data, but I am only presenting the final ones here.

Axial Coding: After I applied open coding to all my data, axial coding encouraged me to focus on the themes, explicitly looking at the various characteristics of each themes. In Table 10 I show some responses to Q10b, which asked how the results of a second sample of 50 spins would compare to the first sample results from Q10a. I've presented the lines of text along with the coded themes.

The main difference between what I did in open coding and what I did in axial coding comes down to focus. In open coding, I collected all responses that had some information "Describing What was Expected". At the same time, I had many memos and tentative themes that I thought might be emerging.

Table 10. *How do you think his results on the second set of 50 spins will compare with the results of his first set?*

Subject	Response	Themes		
		Expected Value	Repeated Values	Range or Extremes
DS	I think it'd be close to it, but different.	●	●	
	So, maybe if he got 28 the first time,			●
	he'd get 24 the second time, or 23...			●
EM	Somewhere near 50%, right	●		
JM	Yeah, I think [it'd be] fairly close in	●		
	the sense that it's gonna be	●		
	around the... 25 blacks,	●		
	plus or minus that 10% or so.			●
RL	I think that it's likely to fall in a			●
	same range, similar range.			●
SP	I think the range would still be			●
	somewhere very similar to that one.			●
	There'd be – just different numbers,		●	
	but still somewhere in that range.			●

In axial coding, I went back through the data and specifically focused on potential themes. For instance, in Table 10, I looked for and coded responses as "Concerning Expected Value" I noticed how DS mentioned that the second sample would "...be close to it", and a comparison of what she had said earlier showed me that what she was suggesting was a result that would be close to the expected value (of 25 blacks). EM and JM also have responses coded at the theme "Concerning Expected Value", with their expression of results being "near 50%" and "around the ...25 blacks". Then, I went back through the data and looked specifically for responses "Concerning Repeated Values". In Table 10, DS and SP specifically mentioned that they expected different results. Similarly, for the theme "Concerning Range or Extremes", JM, RL, and SP explicitly used the term "range", while DS suggested possibilities that range from 23 blacks to 28 blacks.

In the axial coding process, I made many memos to record my own thinking about the data. For example, I noticed when DS mentioned possibilities ranging from 23 to 28 blacks that those were reasonable choices and she didn't seem to expect unlikely extreme values such as 5 or 45 blacks. When RL mentioned range, he first said "same range" but then immediately amended this by saying "similar range". By doing a comparison of other places where RL talks about results being "similar", I was able to find out that in his case and also in the case of most of the students, "similar" seemed to imply "similar but different." By thinking deeply about the themes, I initially recorded my thinking as tentative links between the characteristics of the themes, and the links either became stronger with the addition of more and more data or were discarded.

Constant Comparison: This method lets me take new data and compare it to the themes and characteristics even as those themes and characteristics are emerging. Constant comparison is an ongoing process that can occur alongside and after both open and axial coding. In my previous illustrations of open and axial coding, many instances of constant comparison took place as I referred to other similar survey or interview responses for the different subjects. I also made references to class observations. For instance, in Table 9, JM referred to expecting approximately 50% blacks "plus or minus...20% of that number". Then, in Table 10, he mentioned 25 blacks "plus or minus...10%". By looking at what his group did in class on the posters for the Unknown Mixture, and by referring to my notes on class observations, I could see that a 10% margin of error had been discussed. Thus, JM may have been influenced to think in terms of "plus or minus 10%" even though 20% of 25 is an

easier number to work with. Notice how in Table 9 EM and SP gave ranges of 20 to 30 blacks, which corresponds to 25 blacks plus or minus 20% (of 25). Also, EM had said "40 to 60% of the time" which could be thought of as 50% plus or minus 10%. Constant comparisons let me draw from many data sources as I made different connections about my themes and their characteristics.

Constant comparisons also let me see when I had *saturated* my dimensions, themes, and characteristics. Saturation in this example referred to the point where new data is not adding anything new to the themes I had initially envisioned. Table 11 shows the results for Q10c, which asked for a list of what might happen in six samples of 50 spins each, and why. The same three themes as in Table 10 are again shown in Table 11, which helped confirm that these were themes that I could characterize and look for in other data from other questions. I also was able to compare the subjects' lists for their six samples to the responses they had given earlier. For example, DS had said in Table 9 "within 2 or 3" (of 25 blacks), yet her list reflects 25 plus or minus 4. Notice too how EM had earlier suggested a range of 20 to 30 blacks, but in her own list she went a little wider than that on both sides. JM did a bit of self-reflecting as he considers his range of 21 to 29 and thinks about what he had said about "plus or minus 10%". RL confirmed my ideas about his earlier response when he explicitly stated "similar, but not identical."

Saturation does not imply that one stops thinking of new connections that can be made among the themes and characteristics.

Table 11. *Write a list to describe what might happen in six sets of 50 spins. Why did you choose those numbers?*

Subject	Response	Themes		
		Expected Value	Repeated Values	Range or Extremes
DS	[28, 23, 25, 29, 21, 26] I have a few scattered	●		
	close to 25, but not 25. So that...probably on	●	●	
	average it'll be close to 25.	●		
EM	[20, 23, 28, 32, 18, 24] For the most part, it was			●
	generally concentrated between – give or take...			●
	50% of the time, somewhere near there…But...			●
	occasionally it would be even lower than that,			●
	so I... threw the18 in there...And the 32, yeah.			●
GP	[21, 23, 24, 25, 27, 28] They're around the 25.	●		
JM	[21, 23, 25, 26, 27, 29] Well, they're close to	●		
	that 50 percentile, that we're looking for , plus	●		●
	or minus – I'm thinking – 10% or so. Actually,			●
	I'm a little high aren't I, with the 29? But still...			
RL	[21, 23, 22, 27, 28, 29] I did a pair, each equally			●
	far out from the mean in either direction. And...	●		●
	there's no repeats, but they're all sim[ilar] …		●	
	They're similar, but not identical		●	
SP	[20, 23, 24, 26, 28, 29] Just fall within that			●
	range of 20 to 30, that you would expect. Yeah,		●	●
	they could repeat, but I just did a range acro[ss]		●	●
	from 20 to 30, just to choose.			●

There are so many different layers of meaning that can be found in a grounded theory approach to qualitative analysis that one could continually return to the data and find something new. Thus, being guided by my principle research question to come up with a more descriptive framework for understanding EPSTs' conceptions of variation, I concentrated on describing the characteristics of the dominant themes I saw rising out of the data.

This section provides a taste of what characterizes the three themes for "Describing What is Expected". Concerning the expected value, Tables 9 through 11

show how results are thought of as being "close to", "somewhere near", or "around" that value. Results may or may not include the expected value. For example, in Table 11 when picking six hypothetical results, three of the cases did include 25 blacks in their list, and the other three cases did not. Concerning repeated values, although multiple samples could repeat, they shouldn't all be identical. Results might be "similar, but not identical." Again looking at Table 11, notice how every one of the six cases gave lists that were composed of distinct entries. There is not a single repeated value in any given list. Concerning range or extreme values, subjects sometimes stated explicit numerical ranges which gave clues to the numbers they were comfortable with. In the example of Q10, a range of about 20 to 30 meant that 21 to 29 was considered reasonable, but so too was 18 to 32. Thus, even when stated explicitly, ranges tended to be flexible.

Even with the detailed illustration given within this section, the themes for "Describing What is Expected" (Concerning Expected Value, Repeated Values, and Range or Extreme Values) have not been completely fleshed out just by responses to PostInterview Q10. I have only profiled the six subjects' responses to this question. My purpose in this section has been to show how I used the grounded theory techniques to do a microanalysis of the data. The complete analysis encompassed six cases' responses to the interview questions as well as classwide responses to the survey questions. The illustration of microanalysis in this section relied on less than 50 lines of text comprising subjects' responses in Tables 9 through 11, while the entire research project covered literally thousands of lines of text. Thus, NUD•IST was extremely helpful in managing the codes as they emerged, especially since a key

feature of NUD•IST is the ability assign multiple codes to lines of text. As many of the responses in this section showed, a subject could give a sentence or fragments of sentences that contain several distinct meanings. DS gave a succinct example in Table 9, when she expected results "...close to 50%, but not exactly! Yeah, within 2 or 3". She appeals to the expected value, and also gives a range. In the next chapter I'll summarize each theme for each dimension within the aspects comprising my evolving framework.

Summary

Subjects were chosen from the MET 2 class, where the pedagogical style in the class and in its prerequisite, MET 1, encourages students to communicate their thinking both verbally and in writing. Subjects were familiar with and able to openly share their thinking, a necessary condition for this study. The overall design for the research incorporated elements of both grounded theory and case study traditions. In addition to interview data gathered from six cases, classwide data was gathered via a set of written documents (surveys), and the study was also augmented by in-class observations.

The three data gathering methods, observation, document review, and interviews, all were connected to the three class interventions. The three interventions (made up of two activities each) corresponded to the three contexts for looking at conceptions of variation: variation in data sets, variation in sampling, and variation in probability situations. After administering a PreSurvey to the whole class, PreInterviews were held outside of class time. Then, after each class intervention was

conducted, a take-home PostSurvey was distributed. Finally, PostInterviews were completed.

To analyze the data, the use of grounded theory techniques allowed for the characterization of the subjects' understanding to naturally emerge within the conceptual framework in the shape of distinct but linked themes. The rudimentary structure posited as an initial conceptual framework was validated and extended by doing a microanalysis of all the data collected. The data analysis was facilitated by the use of the NUD•IST computer software. The final result was a rich description and overall characterization of the conceptions of variation held by elementary preservice teachers, given by the evolving framework in the next chapter.

CHAPTER FOUR

Results and Analysis

This chapter has two main sections. In the first section, I present the evolving framework and describe the defining characteristics for the themes within the framework. In the second section, I use the evolving framework to compare the conceptions of variation of six cases from before to after the instructional interventions. The first section addresses my first research question, and the second section addresses my second research question. Both sections highlight tasks that were illustrative in looking at EPSTs conceptions of variation, thereby addressing my third research question.

Evolving Framework

The initial conceptual framework of Chapter Two is organized around three *aspects* of understanding variation (expecting, displaying, and interpreting variation).

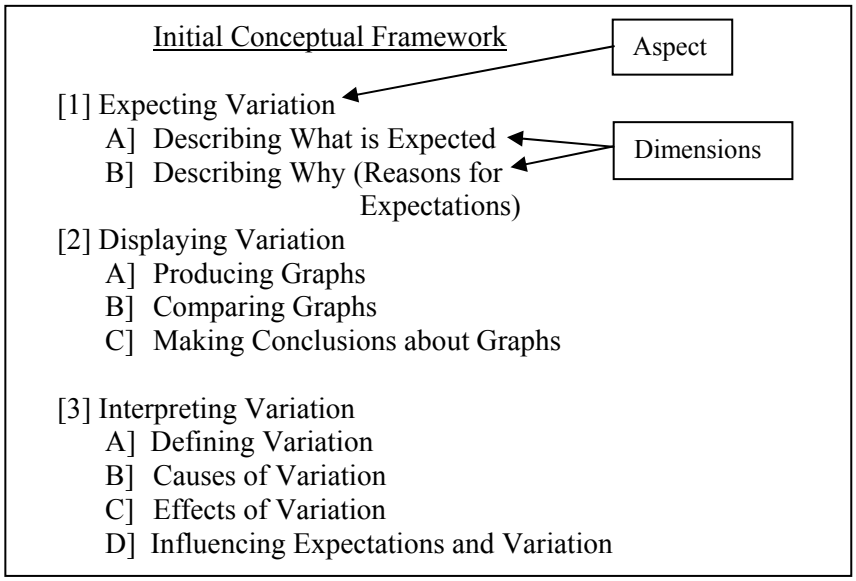


Figure 12 – Initial Conceptual Framework

Each of the three aspects has corresponding *dimensions*, and the aspects and dimensions are illustrated in Figure 12. The last section of Chapter Three showed how the techniques of grounded theory were used to expand the dimensions of the initial conceptual framework into constituent *themes*.

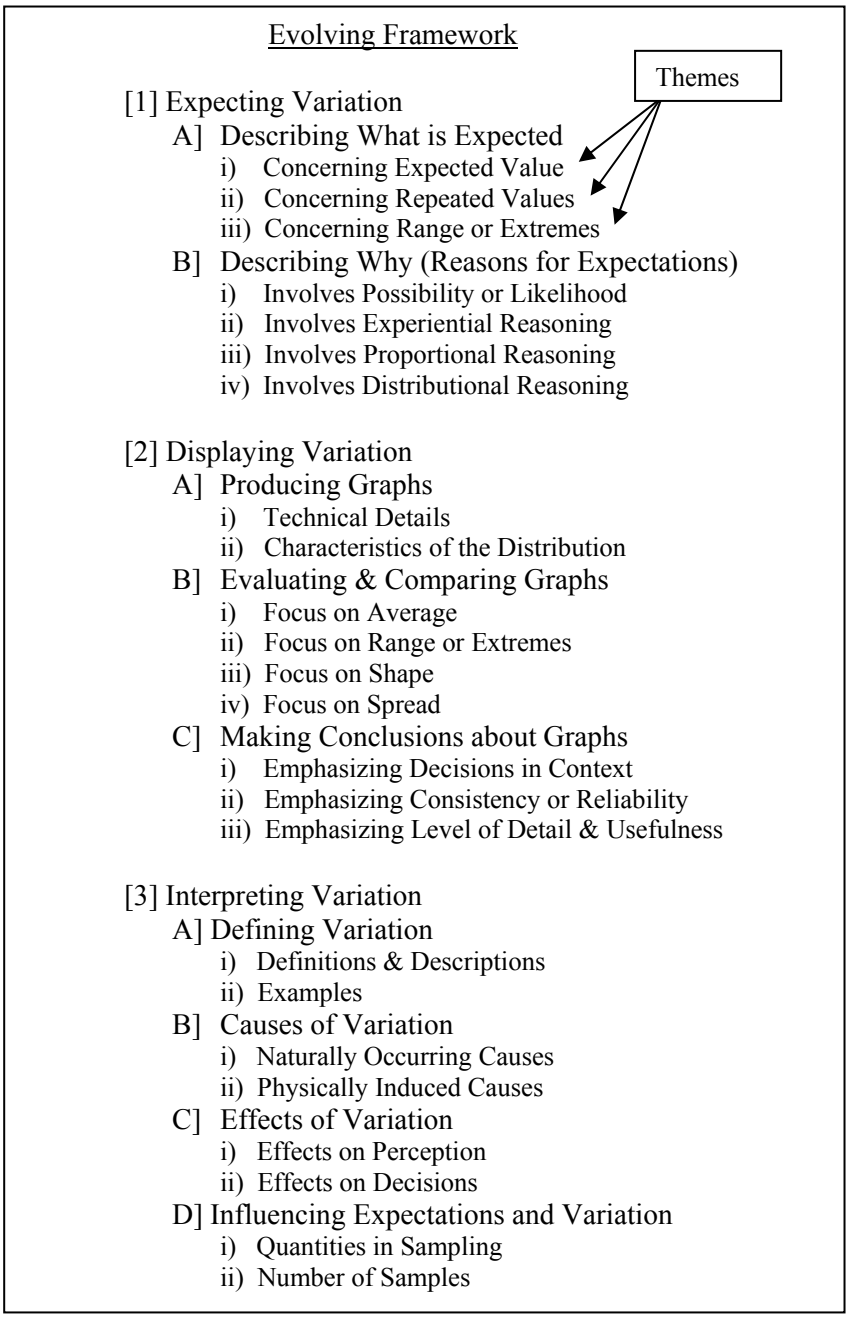


Figure 13 – Evolving Framework

The themes expanded the initial conceptual framework into an *evolving framework* for characterizing EPSTs' thinking about variation (see Figure 13). The previous chapter also showed the method by which a theme was defined by its own *characteristics*. The purpose of this section is to define all of the themes in the evolving framework by describing some of the key characteristics that arose from the data. In keeping with the tradition of a grounded theory approach, the descriptions within this section will be a compilation of my own analytic thoughts stemming from a cumulative consideration of the data, combined with exemplars taken from student responses.

[1] Expecting Variation

A] Describing What is Expected: The three themes within this dimension concern the expected value, repeated values, and the range or extreme values. Although these themes were profiled in the last chapter using data in response to PostInterview Q10, I'll be using data from different questions (mainly from the PreSurvey and PostSurveys which are included in Appendix B) to summarize the themes in this section.

i) Concerning Expected Value – One characteristic of this theme was whether or not results should actually be the expected value. A second characteristic was how results should be close to the expected value, as should the average of results (for example, the average of results from six samples should be close to the expected value for a single sample). A third characteristic was how results should be on both sides of the expected value. I'll illustrate these characteristics next.

Often responses included an explicit or implicit reference to the expected value

for a given situation. For instance, the expected value for the number of heads in a sample of 50 flips of a fair coin in PreSurvey Q7a is 25 heads. An explicit reference to "25" or "25 times" was made by most of the respondents to Q7a in predicting the results for a single sample. An example of an implicit reference is "½ of the time", and further inference may be needed to determine if the subject actually knows what the expected value is. If the response does not include an explicit or implicit reference to the expected value, it is still possible for that response to inform what the subject thinks about the expected value. For example, in Q7a one student put "28" and another student put "24". An interpretation of these results is that these two students knew what the expected value was and yet they chose to put different value. That is, they thought results would *not* be right at the expected value.

Many subjects thought results should be *close* to the expected value. Responses included several different words to convey the same basic idea. For example, in Sampling PostSurvey Q2a, the expected value for the number of red candies when drawing a sample of size 100 from a mixture of 600 red and 400 yellow candies is 60 red. Here is what some students wrote for what they expected in their sample:

BP: Around 60
JM: Very close to 60%
MA: Near 60
SC: Close to 60

Other words or phrases that were used in other questions were "approximately", "about", "relatively close", and "somewhere around". What I learned from responses suggesting results should be close to the expected value is that subjects did have some

sense of variability. A person who puts "Around 60" instead of just "60" is tacitly admitting that he or she would be comfortable with a result that is *not* the expected value, as long as the result is close. As far as learning how close is reasonable for a subject, more information is usually needed.

The responses shared so far for the theme concerning expected value all had to do with predictions for a single result (PreSurvey Q7a and Sampling PostSurvey Q2a). In looking at predictions for multiple results, I found that some subjects also thought the *average* of their predictions should either be or be close to the expected value. In Sampling PostSurvey Q2c, subjects were asked what they thought the results for six samples of size 100 pulled from a population of 600 red and 400 yellow would be. Here is what four students gave for their six predicted results and accompanying explanation:

BP: [56, 60, 60, 60, 61, 63] I chose these numbers because they have a mean, median, and mode of 60. All three are 60.
RL: [50, 60, 60, 65, 70, 75] The mean of the above data is 60
SC: [40, 55, 58, 62, 65, 80] Because they reflect a mean of 60 or 6:4 which is the actual ratio of red to yellow in the container.
SP: [48, 52, 55, 57, 63, 68] If I expect the average to be about 60, then I would guess that the amount chosen would vary above and below 60 & pretty close to 60.

Although RL's choices did not actually average 60, on the basis of his other responses I believe he meant to put numbers averaging 60 and that he was capable of doing so. Predicting multiple results is a good way of getting a sense of what seems reasonable to a subject, and of providing some idea of "how close" to the expected value they really think results should be.

Notice how BP and RL actually included the expected value of 60 in their lists,

while SC and SP did not. However, a commonality in all four lists is that results are given on both sides of the expected value. Having multiple results that are both higher and lower than the expected value was a feature of many of responses. In other words, responses seldom suggested that results be only on one side of the expected value. Subjects sometimes explicitly explained their choices by stating how the chosen numbers were on both sides of the expected value. In Probability PostSurvey Q1c (with an expected value of 25 blacks for a sample of 50 spins of the fair white-and-black spinner), here are three responses that predict the results of six samples:

DS: [18, 21, 24, 26, 28, 31] None of the #'s are too high or too low (far from the 25) which would be hard to hit based on the 50% odds
EM: [20, 22, 24, 26, 28, 30] Because they are all near to the 50% of the time landing on black. Sometimes above and sometimes below.
SP: [22, 23, 24, 25, 26, 27] They are a bit higher, & a bit lower, or are 25... Which is the expected ratio?

The above three subjects specifically called attention to results on both sides of the expected value.

There were other characteristics for this theme (such as how often the expected value might repeat, or how far away from the expected value results might be), and I've only chosen to illustrate a few of the dominant characteristics. The chief commonality was that in order for a response to reflect this theme, the response must tell the reader something about what the subject thinks in relation to the expected value.

ii) Concerning Repeated Values – This theme was reflected in responses to questions involving multiple samples. The main characteristics of this theme shared

the common concern of whether or not results from multiple samples should repeat, and if so then how much repetition was reasonable to expect.

A characteristic on one end of the spectrum was that multiple results should all be the same. That is, repetitions were seen as likely. For example, in PreSurvey Q1b, RL said "Yes" when asked if he thought the same result would occur for several samples. In PreSurvey Q1c, when asked to predict the results for six samples, two of the 25 students' responses were "6, 6, 6, 6, 6, 6", which told me that those two subjects expected homogeneous results for the six samples.

At the other end of the spectrum, some subjects thought a list of results should hold no repetitions. Again citing an example from PreSurvey Q1c, two responses showing the characteristic of no repetitions were "3, 4, 5, 6, 7, 8" and "2, 3, 4, 6, 8, 9". The responses of DS, EM, and SP for Probability PostSurvey Q1c also reflect the characteristic of having no repetitions. Some students wrote explanations about how results should be different every time, such as SZ in response to Sampling PostSurvey Q2b: "Lots of possibilities. Won't get same number each time. Random selection." Another student, SR, wrote for Q2b that "every pick is different."

In between the ends of the spectrum, most responses for this theme had the characteristic of allowing for some repetition and some differences among multiple results. In their explanations, subjects tended to either emphasize how results would be similar to or different from one another, but sometimes responses conflated both similarities and differences. I'll cite some examples from PreSurvey Q7b, which asked subjects to predict how results from a second sample would compare to the first.

BP: I think results will be close, but not exactly the same
JX: Similar, though probably a little different
SP: Could be similar, could be completely different
SX: They will be similar but not the same
TO: May vary a little, but not much

Notice how TO used the word "vary" to imply that there will be differences in repeated results, but she qualified her prediction and suggested that results will not differ by much.

Thus, responses for this theme all address how much or little multiple results should repeat. The following three responses nicely illustrate the characteristics of all results repeating, all results being different, and some results repeating. The examples are taken from the Sampling PostSurvey Q2c:

RB: [60, 60, 60, 60, 60, 60] I chose 60 for each classmate because theoretically you should always get 60 red and 40 yellow. This would be the most educated guess at the 6 outcomes
SX: [50, 58, 60, 61, 62, 66] They are all numbers close to 60 but all different to account for variation.
LW: [52, 58, 59, 60, 60, 62] I still believe 60 would be pulled most often. If not sixty then a number close to that.

Each of the above responses reflected a different perspective on the theme concerning repeated values, and each response said something about how the corresponding subject viewed variation in the context of the problem. RB saw no variation as "the most educated guess", but in fact his response demonstrated a naive view of the variation that would likely result. For SX, the presence of variation means that all six samples should be different, which is not an unreasonable expectation for this problem. Lastly, LW's response appealed to the valid idea that 60 is the mode for the underlying distribution, but for any set of six samples it is not very likely that the

expected value would occur twice.

iii) Concerning Range or Extreme Values – Responses for this theme expressed what was expected in terms of a range. Sometimes subjects stated an explicit range (using numbers), and sometime they just referred to expecting a "range" of results. I also phrased this theme in terms of extreme values because occasionally a response seemed to focus on only one end of the range or the other. In illustrating some characteristics of this theme, I'm including examples only from questions that did not explicitly ask for a range answer. That is, I'm choosing questions for which the subjects volunteered range answers or comments about the extremes on their own accord.

The Sampling PostSurvey Q2b elicited some responses showing a vague form of range expectation in the sense that a range is implied or stated, but we don't see exactly what range the subject might expect or allow. Some sample responses are as follows:

MG: It will be relatively close to 60, but the number will vary, sometimes you'll get more red and some less
SA: I think that the mean would be around 60 (for reds), but there would also be other numbers higher and lower than 60
RL: Taking samples, I expect to see a RANGE of data, not specific values.
SP: If you repeated this many, many times, the average would come out to somewhere around 60, but there would be a range of # of reds because you are choosing randomly.

MG and SA's responses both implied that those two subjects would be comfortable with results being within a range, and RL and SP specifically mentioned the word "range." More explicit range responses for Q2b include BP's comment that "there is a

CHANCE to pull out anywhere from 0 to 100 reds". BP's range is unreasonable but numerically explicit. JL also gave an explicit numeric range, which was fairly reasonable: "I believe if you pull and remix, you will get 58-64 reds every time, give or take 60% reds." I was unclear about what the "give or take 60% reds" meant to her, but I thought JL's expectation of a range from 58 to 64 reds was quite appropriate to the situation. It was more common to get numeric ranges as responses in the Probability PostSurvey, suggesting an increased appreciation for how results vary. In Q1b subjects were asked to predict how results from a second sample would compare to the first. Here are a few sample responses:

GP: It could be 28 or 20 or 16
MA: This will give him a score between 24 and 27 black.
MM: Maybe a little different but still somewhere around 20-30
SC: Maybe a little wider range 18 – 32

All of the examples for this theme so far appeal to both sides of the range without much emphasis on one side over the other. Sometimes responses explicitly contained information that told more about what the subject expected regarding one end of the distribution. In Sampling PostSurvey Q4a, subjects were asked to predict 50 results for pulling samples of size 100 from a population 600 Red/ 400 Yellow candies. The subjects assigned frequencies to bins of 0-10 reds, 11-20 reds, on up through 91-100 reds. To explain their choices on Q4b, here is what two subjects wrote:

CS: Still going for the odds of 60 / 40. Most I think would be between 40-80. The 81-90 I chose three.
JX: Because the highest single amount would be 60 reds, since the ratio is 600/400. Then the next higher amounts would be on either side of that, and decreasing out both ways with just a few at the lower #s.

Notice how CS emphasized the upper extremes, with a few results in the 81-90 bin (such results would actually be fairly unlikely in the context of the question). JX, on the other hand, emphasized the lower extremes, and she had listed one result in the 11-20 bin, which also is unlikely. The point is that these subjects gave additional information about only one end of the range.

Another way in which subjects commented on extreme values was when they judged graphs as real or made-up. To illustrate, I'll use PostInterview Q5, which included two graphs purportedly showing results from two different classes. Class A made 40 samples of size 10 from a population of 60 Red / 40 Yellow, and Class B made 40 samples if size 100 from a population of 600 Red / 400 Yellow. In expressing her doubts about Class A's results being real, EM said: "Well, you know, with 40 pulls it seems a little less likely that you would have some on the lower end , you know...". Talking next about Class B, she went on to say:

> EM: And, actually, I would say the same thing for Class B, 40 pulls but 100 pieces, I would expect between 50 and 80 to be pulls but 100 pieces, I would expect between 50 and 80 to be where it is here, and then 81 and 90, I would definitely think that , that seems alright to me, but there's two or three that pulled between 21 and 30, and that seems a little low...

For both Class A and Class B, EM felt that there were too many low extremes.

B] Describing Why (Reasons for Expectation): Responses for this dimension addressed any of four themes involving possibilities and likelihood, proportional reasoning, experiential reasoning, and distributional reasoning.

i) Involving Possibilities and Likelihood – With this theme, *what* subjects expected often came alongside a reason for *why*. For example, some subjects expected

to see a sample result of 60 reds from the Large Jar (containing 600 Red / 400 Yellow) because 60 reds was the most likely result on any given trial. Repeated results were unexpected because they were seen as unlikely. Extreme values were often described as unlikely but possible. Included in this theme were responses characterized by similarly vague language such as what might or could happen. Subjects also used probabilistic language in a general way, talking for instance of how the chances for events were seen as high or low. The subjectivity for the class of responses within this theme could be also seen by the way students often would stress their impressions of outcomes, using phrases such as "highly unlikely" or "very possible."

In PreSurvey Q7a, RL predicted "25" heads for the result of 50 flips of a fair coin. SX predicted "24". Both subjects used the language of likelihoods in explaining their respective choices:

> RL: [25] It's the most likely scenario; there's no reason to believe (i.e. no external force) that either of the two outcomes will appear more often.
> SX: [24] It will be close to 25 (1/2 of 50 = 20 x 2 sides) but the likelihood that it lands on 25 is small.

What is interesting is that RL sees the expected value of 25 as "most likely" while SX see the likelihood of attaining that value as small, and both perspectives are reasonable for the context of the problem.

As is typical for responses within this theme, there are no quantitative clues in the above responses for just how likely or unlikely the outcome is perceived by subjects to be. Similarly, in PreSurvey Q1a, DS claimed that for one sample of size 10 "I MIGHT get 6 red". Then she added, "Although it's possible to get 10 yellow." I didn't get a sense of just how possible 10 yellows seemed to DS, and a this lack of

clarity was echoed by other responses about what could happen. Consider the following responses for PreSurvey Q1b about comparing several sample results:

CM: One might return with different combinations
SP: Each session could produce different results
JL: The likelihood that every grab yields 60% reds is just not there
SC: It can't possibly always be the same

Notice how the first two responses are phrased in terms of what is possible, while the the last responses are cast in terms of what is not possible.

I found that Sampling PostSurvey Q1b was very helpful in gaining data on how likely an extreme outcome was perceived to be. The language in the responses still contained a great deal of language along the theme of possibilities and likelihoods, but a part of the question required a numeric prediction that helped me better understand their explanation. Q1b asked how many samples of size 10 would need to be drawn from the Small Jar (60 Red / 40 Yellow) in order to achieve a result of 0 red or 1 red. Two subjects explicitly concluded that such results were not possible:

RB: There is no amount of [samples] that will guarantee no or one red candy in a sample
SA: It's impossible to get zero red. It would take hundreds of tries to get just one red.

Here are two other responses that talk in terms of likelihoods, but for this question I was able to refer back to the number of samples they had predicted (which I have listed below in parentheses):

LW: [500 samples] The odds are very unlikely that someone will pull 0 or 1 red candy
SX: [Thousands] Because the likelihood is so small that only 0 or 1 red candy would be pulled.

For LW, the 0 or 1 result was "unlikely", which translated into needing 500 samples before such a result occurred. SX equated "small" likelihood in the situation with a need for thousands of samples, but she didn't say how many thousands. The actual expected value in the context of the problem is close to 10,000 samples.

This theme encompasses all the responses that included subjective language about what was possible or likely, and what could or might happen. I found that every subject did at some point use language reflecting this theme, and I think the reason is because subjective probabilistic language is part of our natural way of speaking. In an interview setting, I found it useful to ask for examples when a subject started using unclear probabilistic language. For example, if someone said an outcome had a "high probability" then I would ask how high. In a written survey context, questions that ask for specific examples help define what a person means when they use language of possibilities and likelihoods. However, saying that an outcome is impossible or could not happen is clear enough.

ii) Involves Experiential Reasoning – The two characteristics for this theme are informal and formal experience, with the commonality that both characteristics appeal to having previously seen or done or heard about a similar situation. Informal experience includes time spent playing games at home, for example, whereas doing a class activity involving game playing is classified as formal experience. Responses like "I usually roll 6's" in reference to dice expectations are classified as informal experience. A response such as "From what we did in class, I know that 6's don't happen that often" would be based on formal experience. I made the distinction between informal and formal experience as a way to group the responses I found for

this theme.

To give some examples of informal experience, I'll use a Q1bii from the Data & Graphs PostSurvey. Q1bii asked whether subjects thought Portland (Oregon) or Columbus (Ohio) was rainier, and why. The question came on the heels of earlier questions showing graphs of 30-year averages for monthly rainfall in the two cities. Here are three responses:

MM: I think Portland is rainier from personal experience and general knowledge of Columbus Ohio.
SR: I have spent time in both places and Portland is rainier.
SA: Portland because I live here and it rains all of the time.

All three of the above subjects have phrased their responses so that personal opinion dominates their reasoning. The responses reflect informal experience in the sense that the subjects merely stated what their sense of the situation was, absent of any formal data analysis.

Regarding formal experience, I thought mainly of information gleaned from structured classroom activities. For my research, I was interested in hearing if the subjects would mention the impact of the interventions done in class. With the Sampling PostSurvey Q1b that concerned how many samples were needed to get an extreme result of 0 or 1 red candy, many subjects commented on the classroom intervention we had done on sampling. A particular impression was made by the computer simulation using ProbSim that we did as a class, whereby we had a class discussion even as Matt continually ran the simulation with more and more samples. Here were some of the reasons offered by subjects for their predictions on the Sampling PostSurvey:

DP: When we did over 5000 tests via the software program, we STILL didn't get the lower #. Chances are very SLIM
EM: After seeing the simulations in class on the computer, it seemed almost impossible to get a zero.
MG: When we did a similar exercise in class, we were only able to do it with a huge number of attempts.
SA: I know this because we saw it on the computer program in class.
SL: I based it on the activities we have done in class w/ computer program as well as hands-on activities where we never got 0 or 1
SP: I was thinking about the simulation in class and how many trials we had to enter in the computer until we got a 1

I've included more than a few sample responses for Q1b to emphasize the impression that formal experience made on the subjects. The main effect on the subjects seemed to be more in their reasoning than in their actual predictions. For example, DP remembered that it really did take "over 5000" samples to get the extreme results of 0 or 1 red. However, most of the other subjects couldn't recall if the numbers were in the hundreds or thousands, just that it took (as MG put it) "a huge number of attempts." To help subjects make more realistic expectations for *what* might occur, I recommend more use of class activities and computer simulation. As far as influencing subjects' reasoning *why*, it was clear that even the twenty minutes computer simulation that we had done in class (which followed an activity of hand-drawing the samples) made a lasting impression on the subjects.

iii) Involves Proportional Reasoning – Proportional reasoning was a part of almost every student's explanation at some point in their individual responses. The variety of ways they collectively had to explain included ratios, decimals, odds, and fractions. Since MET 1 was a prerequisite for MET 2, and proportional reasoning receives a fairly in-depth treatment in MET 1, I had expected the students to be able to reason proportionally. Some sample responses from PreSurvey Q1a (reasons for the

predicted results for one sample of size 10) show the diversity that subjects had in expressing their proportional reasoning:

DS: Because 60% are red so odds are I'd get 6
DM: Because the ratio is 60 Red: 40 Yellow out of 100, so when you grab 10, the likelihood of the ratio being 6:4 is high
BP: Because 3/5 of the candies are red, and 3/5 of 10 is 6
SA: Because is you have 100 candies and 60 are red, when you have 1/10 of that, 1/10 will still be red
SC: Because it is the most probable amount since the ratio 6:4 exists throughout the container

Proportional reasoning was fairly easy to identify in subjects' reasoning, and for some subjects such reasoning was a dominant strategy. An interesting example of non-proportional reasoning came on the PreSurvey Q5b, which asked for a comparison between two classes' test results (the two class sizes were different, but they had taken identical tests). Most of the subjects misidentified the class that had better test results, and one student wrote that a comparison "cannot be determined since the classes held different numbers of students." I found it curious that my subjects, almost all of whom could reason proportionally on questions about sampling and probability, were not as quick to apply ratio thinking in PreSurvey Q5b. I think one reason is because of most students' poor ability to reason with data and graphs.

However, most student responses involved proportional reasoning to different degrees on different tasks having to do with sampling and probability, and typically the *why* of proportional reasoning went together with an average for *what* was expected.

iv) Involves Distributional Reasoning – Of all themes involving reasons for expectations, the theme involving distributional reasoning is what best encompasses a

richer appreciation of variation. Reasoning about possibilities and likelihoods, or arguing on the basis of experience, or using proportional reasoning can all contribute to a better understanding of variation, but distributional reasoning really lies at the heart of this research. I'll describe what I mean by distributional reasoning, and then present some examples focusing on the characteristics that I looked for in this theme.

Distributional reasoning involves a consideration of the distribution of a set of data, or the distribution underlying a situation. For example, in predicting the results for a single sample of size ten from the Small Jar (as in PreSurvey Q1), it is helpful to consider what the sampling distribution for many samples might look like. Features of a distribution include the center, or average, but distributional reasoning goes further than just a consideration of center. Other important features of a distribution include the range, shape, and spread of the distribution (Shaughnessy, Ciancetta, Best, & Canada, 2004).

Each of the features of center, range, shape, and spread are themselves multi-faceted. Centers can be thought of in terms of mean, median, and mode. Ranges might be considered as the maximal minus the minimal values, or a trimmed mean might be of interest. Shapes are often described in terms of their visual characteristics, such as flat, bell, skewed, or bimodal. I distinguish spread from range to emphasize the way that data clusters close to a center, or spread from the mean, or is concentrated at various intervals within the range.

Distributional reasoning can therefore encompass several different characteristics, and some subjects incorporated elements of distributional reasoning into their responses to varying degrees. I'll discuss some of the better responses

exemplifying this important theme, choosing examples from the PreSurvey and each of the PostSurveys. In PreSurvey Q2, students were asked to predict a range for six samples, then a range for thirty samples, and then to offer an explanation. MA listed a range of 3 to 9 for both six and for thirty samples, and wrote

> MA: I chose a wide range of red candies to begin with. I feel it is more likely that this range will happen when more people do the experiment. However, there will be a greater grouping near the six red candies than any other number

What I liked in MA's answer was the language about a "greater grouping near the six red," which I felt demonstrated an understanding of the spread in this situation. In explaining her choices for predicting 50 samples on Q3 of the PreSurvey, DS rationalized that

> DS: Most people would be close to the 60% of total # of reds. Fewer people would be at the far ends of the curve (a lot higher or a lot lower than 60%)

Notice how DS incorporated both a sense of clustering around the average as well as a sense of paucity of data at the extremes. In Q5a, there were graphs portraying the test results of two classes of equal sizes. The graphs were both symmetric and had the same means but different ranges. SC wrote a very detailed response in evaluating the two classes:

> SC: The two classes taken as a whole did equally well on the test, with an average score of 5, but individually there was a student who scored lower in the Brown class – But this was offset by the higher score in the same class for another student. I could see it was the same by see the symmetrical arrangements of the shaded boxes – Both with 5 being the "highest" on the graph – Simply restacking the boxes – putting on either side (3 & 7) – Turns the yellow class into the equivalent of the Brown class.

SC wrote about the average and how data was distributed about that value. She also

mentioned the shapes of the distributions, and in describing her "restacking" she exhibited a powerful sense of distributional reasoning for this question.

In the Data & Graphs PostSurvey, Q2a presented graphs for rates of traffic fatalities between the South and the Northeast regions of America. Here are three responses that I offer because of the distributional language that they contain:

EM: The highest concentration of data for the North is clustered at the "lower" end around 1.5 deaths, while the concentration of data in the south is decidedly clustered to the high end above 2.5 deaths

MA: The median deaths for the South = 2.6 , for the Northeast = 1.6, the mean for the South = 2.46 where for the Northeast it's only 1.79. Also the concentration of deaths is more compact around the interquartile [range] for the Northeast.

SC: It seems like there are more traffic deaths in the South than the Northeast from the data I see on the dot plot and boxplot. The median is higher and the concentration higher for the South

The three subjects above all conveyed a sense of distribution through terms like "concentration" and "clustered" in reference to the spread of the data. Other terms like "grouped" or "clumped" or "bunched" can help describe how data gets distributed.

When justifying predictions for 50 samples of size 100 taken from the Large Jar on Sampling PostSurvey Q4b, again we see distributional reasoning in the following four responses:

JB: The highest number of people draw from 51-60. If graphed, the graph would be symmetrical.

MM: Seems like there would be a concentration here [in subrange 41-70] and then the others would be the outliers or less likely pulls

SC: Because they create a "picture" of data that peaks around 60 and clusters around that mark, diminishing as it moves to the extremes of 100 and 0.

SA: Because 51-70 is closest to the mean, so those will happen the most times. As the numbers get farther away from the mean, they will happen less.

Recall that the bin widths (such as 00-10, 11-20, etc.) were given in the problem statement and did not originate within the subjects. However, notice how all the above four responses indicated a clustering of data near the mean, and the last three specifically implied less data farther away from the mean.

As a final set of examples to illustrate the theme of distributional reasoning, consider these reasons that subjects gave for their predictions of six samples of the fair spinner on Probability PostSurvey Q1c:

RB: I would expect the outcomes to fall into a bell-shaped curve much like this: [He has drawn a bell curve centered at 25]
RL: These numbers represent a distribution across a range of likely results
SL: The spins would probably be concentrated in the central to upper 75% range since that seems to me the way the data usually goes, but the numbers were random.

In SL's response, note how she called attention to a subrange within which she expected data to be "concentrated".

I gave more examples for this theme than for previous themes because of its importance in revealing thinking about variation. The individual characteristics within the theme – centers, range, shape, and spread – are at least as important as the mix of these characteristics within a response and the language that subjects use. Many of the examples given for this theme are lengthier precisely because some subjects were relating different elements of the distribution together. Subjects also may lack conventional terms such as "standard deviation", but they are still capable of conveying a sense of reasoning about the distribution of data.

[2] Displaying Variation

A] Producing Graphs: This dimension addresses the questions that asked students to draw their own graphs to predict outcomes for situations in sampling, data-driven, and probability situations. The two themes which stood out to me in the kinds of responses were technical details of the graphs and also the characteristics of the distribution shown in the graphs.

i) Technical Details – This theme has to do with a subject's graph sense. Characteristics of this theme included the type of graph the subjects used and also the appropriateness of the scales and labels along the axes. To achieve the goal of getting students to illustrate in their graphs the kind of variation they expect in a situation, the students need to have command over the type of graph chosen and also a sense of how choices of scale along the axes can affect the appearance of variability.

Types of graphs that subjects used included smooth curves, bar charts, dot plots, scatterplots, pictographs, and straight lines. Sometimes they labeled their axes and put appropriate scales and sometimes they did not. To illustrate some of the different types of graphs, first consider PreSurvey Q4, which asked for a graph showing predicted results of 50 samples of size 10 taken from the Small Jar. GP drew the skewed bell shown in Figure 14. I had provided labeled axes on the PreSurvey, and placed a scale on the horizontal axis. In GP's case and several other subjects, no scale was given for the vertical axis, making it hard to tell how plausible his graph was.

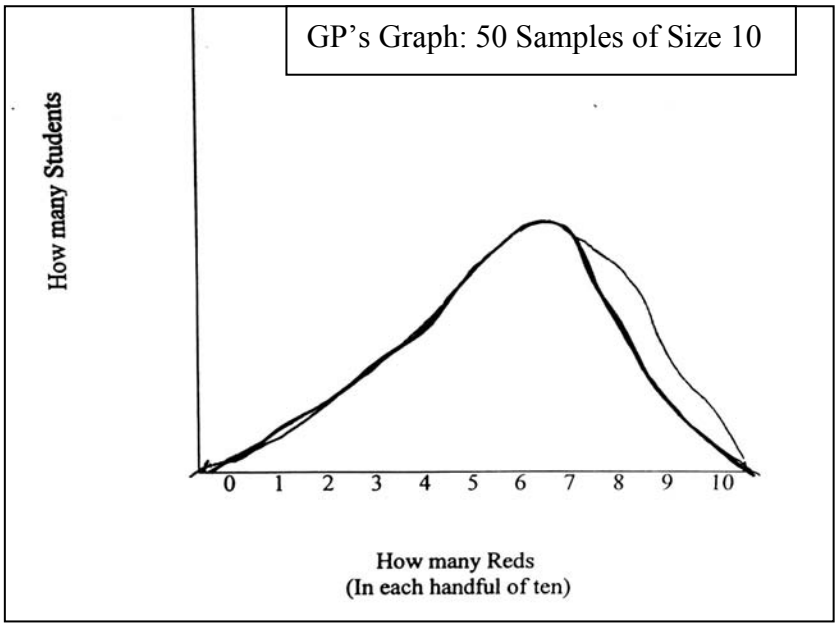


Figure 14 – GP's Response to PreSurvey Q4

I also found it curious that so many subjects used continuous curves for PreSurvey Q4, since the sampling experiment had only 11 possible outcomes (0 Reds through 10 Reds). By the end of the research there were many more types of graphs used. I think the reason so many people used smooth bell-shaped curves in the PreSurvey is because some previous class experience has impressed upon them the significance of such curves. I suspect that the probabilistic heuristic of *availability* has a counterpart in statistics, and when it comes to graphing predicted outcomes older students (such as my research subjects) automatically think of a smooth bell curve.

By the time of the PostSurveys, we had practiced making several different types of graphs in class. I'll next share a response to Data & Graphs PostSurvey Q1c. The question asked for a graph showing how many inches of rain Columbus, Ohio might get for each day in June, assuming that the average monthly rainfall for June

was 4 inches. Figure 15 shows BP's graph. Again, I had pre-labeled both axes, and I also had subdivided the horizontal axis to show marks for each day.

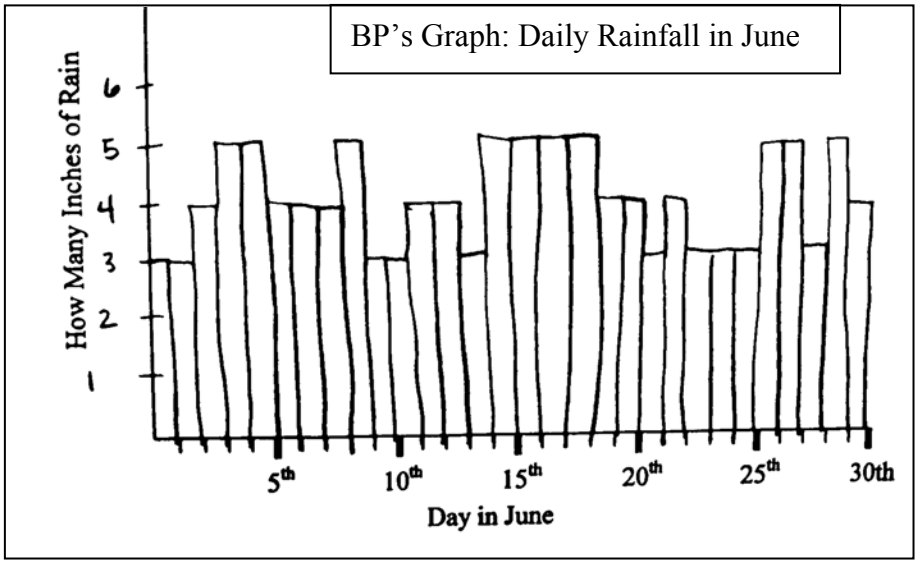


Figure 15 – BP's Response to Data & Graphs PostSurvey Q1c

In BP's bar chart, she showed an incorrect vertical scale. The idea that 4 inches is the average monthly rainfall for June means that a daily average could be thought of as (4 inches) / (30 days) = 0.13 inches per day, with variation. BP's graph erroneously implied a *daily* average of 4 inches, not a *monthly* average.

On Probability PostSurvey Q3, subjects were asked to graph predicted results for 40 samples of size 50 from the fair spinner. I had labeled both axes, but only scaled the horizontal axis by five. JL's graph is shown in Figure 16.

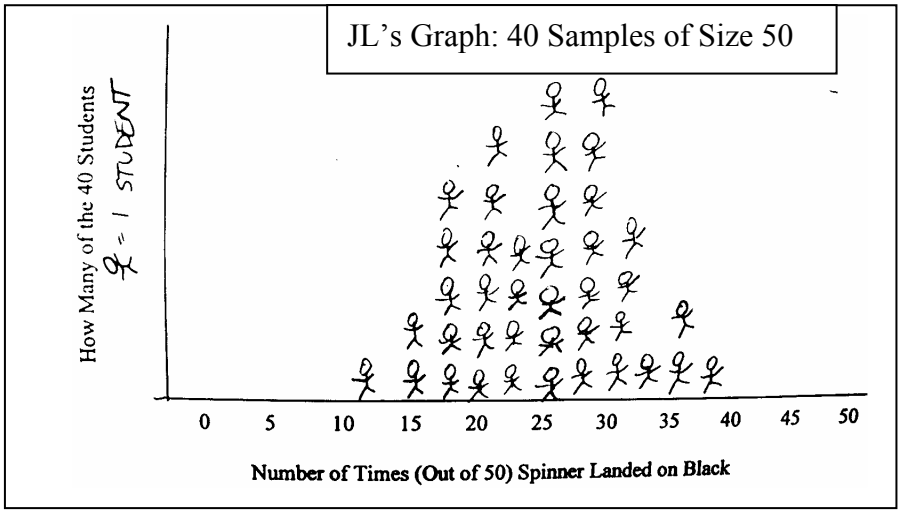


Figure 16 – JL's Response to Probability PostSurvey Q3

Using a pictograph, JL didn't need a scale on the vertical axis, but did need a legend (which she provides). Along the horizontal axis, it isn't quite clear what value each of her columns of stickmen are aligned at, but I did count and see that she drew every one of the 40 required stickmen for her graph.

It seems clear that the way in which students produce graphs to show expected variation depends not only on their own sense of variation but also on their repertoire of different graph types and their skill in conveying necessary information on the graph (via proper use of axes, for example), in other words, technical details.

ii) Characteristics of the Distribution - When the technical details of a graph are plausible or at least understandable, the characteristics of the distribution can be assessed. The four characteristics that I found most salient to understanding variation in EPST's graphs corresponded to the same four characteristics for the theme of distributional reasoning. Those four characteristics concerned the center, range, shape, and spread of the distribution. For example, the center (or average) may be too

high or low, or the range may be too narrow or too wide. Shapes of distributions can vary, particularly in sampling or probability situations involving small amounts of data. Spreads may be too tight or too scattered, or the data may look as if it is unnaturally distributed.

I'll give some new examples from the same three questions that I profiled in the previous theme, since those questions were the only ones from the Surveys in which subjects were asked to produce graphs. In PreSurvey Q4, GP's graph (Fig. 14) had the shape of a skewed bell. SP's graph for the same question is shown in Figure 17. Note how SP included a scale on her vertical axis, and she had also placed some points on the graph to show the frequency for each outcome.

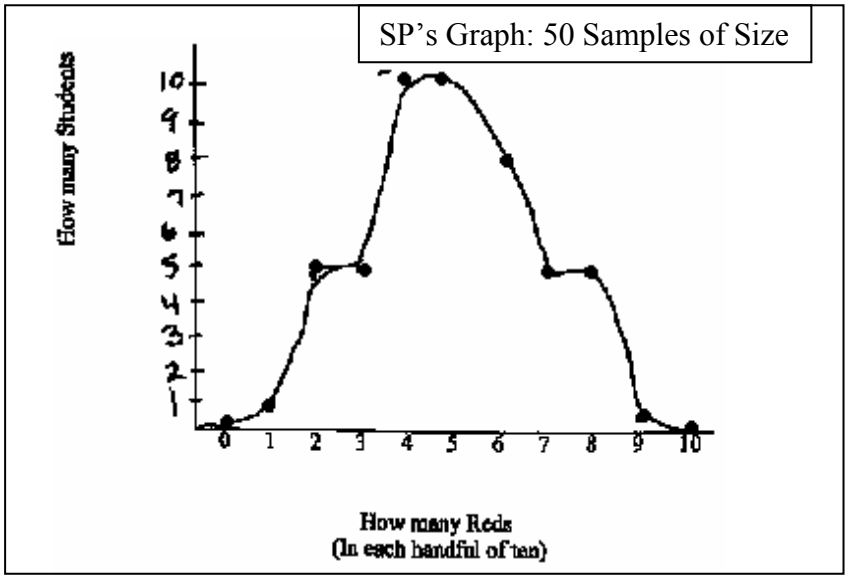


Figure 17 – SP's Response to PreSurvey Q4

There are two immediately questionable features of SP's distribution: First, her graph had a symmetrical shape, and second, it is centered around 4 and 5 Reds (which corresponds with her claims that results would be in the "midrange"). I also think she had an unrealistically low expectation concerning the upper end of the distribution.

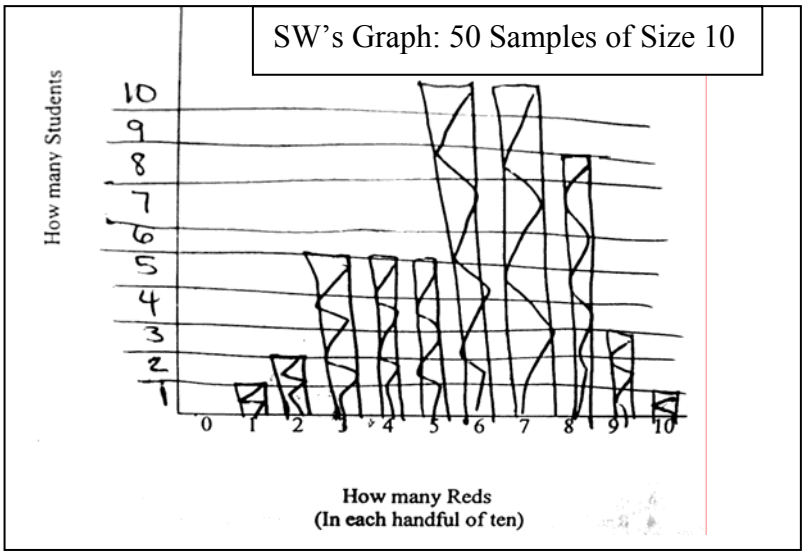


Figure 18 – SW's Response to PreSurvey Q4

In contrast to SP's graph, SW used a bar chart which had a more realistic center and range, and which showed a reasonable shape but a wide spread (see Figure 18).

The average rainfall graph asked for in Data & Graphs PostSurvey Q1c elicited some interesting interpretations of what subjects thought might be a reasonable shape for the distribution. Some subjects had the inches of rain going up and down every other day, while others had no rain for several days followed by some rain for a few days. In BP's graph shown earlier (Figure 15), aside from the incorrect center around 4 inches of rain per day, she also had it raining every single day in June. In contrast, MA's line graph showed most days as having no rain (see Figure 19). Although MA's scale on the vertical axis is coarse, it is easy to guess that her chosen values for days with rain are (from left to right): 0.5", 0.5", 1.5", 0.5", and 1.0". Her range is realistic, going from 0" to 1.5", and her graph implies an average of 0.13 Inches Per Day = (4 Inches)/(30 Days) , as expected. Choosing convenient numbers that easily add up to 4" was common for many students who obtained the correct average.

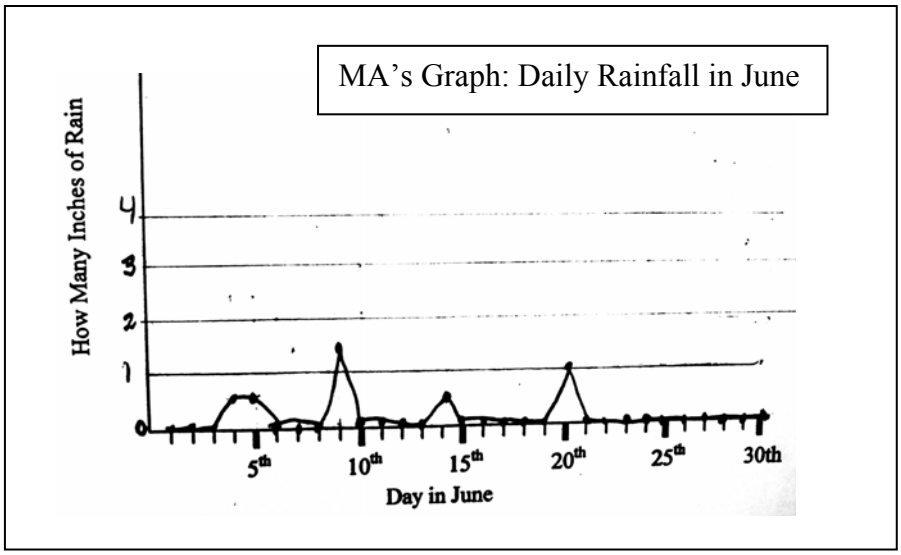


Figure 19 – MA's Response to Data & Graphs PostSurvey Q1c

Some students, however, had a correct daily average of 0.13 Inches but somehow missed the point of variation in weather patterns. RB gave a good example of uniform distribution in Figure 20.

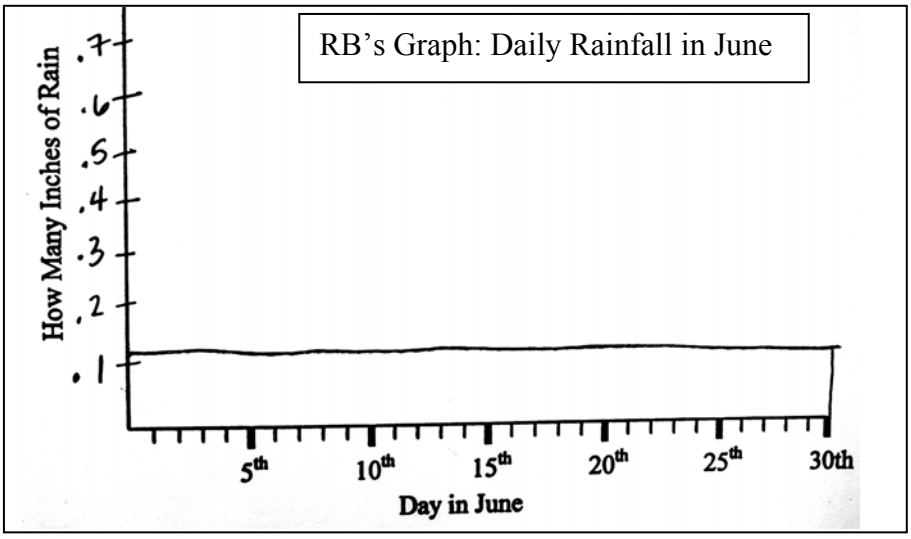


Figure 20 – RB's Response to Data & Graphs PostSurvey Q1c

RB's horizontal line conveyed the notion of absolutely no variation, and a steady flow

of rain all through the month of June. Another subject, RL, also used a straight line but had the line increasing from left to right, saying that the "graph is inclined because [the] average for July is greater than 4". Most subjects showed some day-to-day variation in their rainfall distributions, even though most did not correctly determine the daily average.

As a final example, consider BP's graph for Probability PostSurvey Q3 shown in Figure 21. Whereas JL used a pictograph for this question (shown earlier in Figure 17), BP decided to use a bar chart.

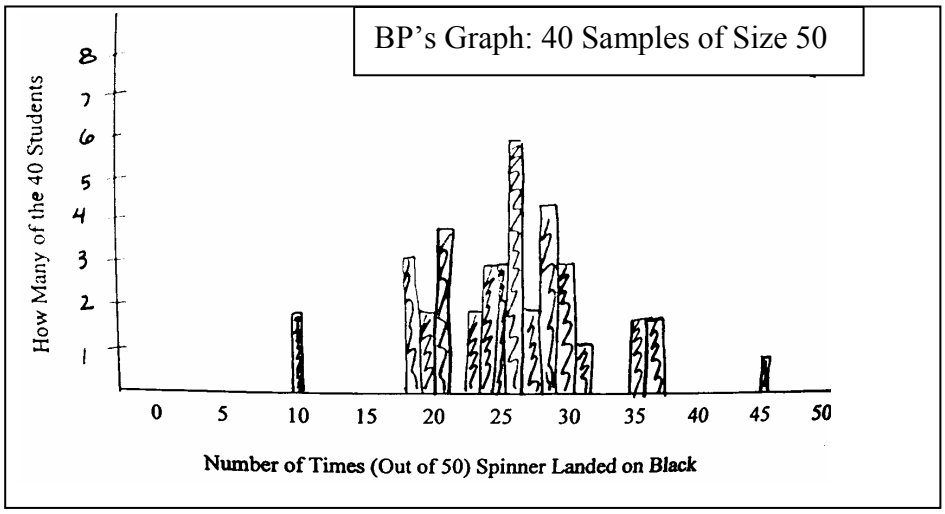


Figure 21 – BP's Response to Probability PostSurvey Q3

BP's center was reasonable, but her range went a bit too high. She had a reasonable clustering near the expected value of 25 blacks, and also had uneven staggering of the bar heights. By uneven staggering, I mean that frequencies are not uniformly increasing or uniformly decreasing on either side of the mode. JL's graph (Figure 16) also showed an uneven distribution, which contrasted with graphs that tapered off from the mode on both sides.

Thus, in assessing graphs that subjects produced, I considered both themes concerning *technical details* and *characteristics of the distribution*. Whereas on the PreSurvey I mostly saw smooth bell-shaped curves, as the class progressed the subjects gained more fluency with different graph types and they improved in their attention to technical details in making graphs. Also, on the whole I saw more graphs towards the end of the research which displayed reasonable centers, ranges, shapes and spread of the data.

B] Evaluating and Comparing Graphs: The four themes that I included for this dimension corresponded to the four characteristics of distributional reasoning described earlier. One way to look at this dimension is that I have amplified my attention to the four characteristics, considering responses in terms of how they focused on the average, range or extremes, shape, and spread of data as shown in graphs. Distributional reasoning is an important theme in reasoning about expectation, and characteristics of the distribution are therefore important when subjects produce their own graphs. I wanted to see if they attended to these same characteristics of the distribution when evaluating and comparing graphs. What follows next are specific examples of how subjects referred to averages, range or extremes, shape, and spread of data when they were evaluating or comparing graphs.

i) Focus on Average – On many of the tasks, a box of summary statistics was provided, listing the mean, median and mode for the data. The box was usually put in close proximity to the graphs, and some subjects were so influenced by those numbers that they admitted to not even paying much attention to the graphs. All they needed to know was what the measures of center were.

On the PreSurvey, I did not provide any of these summary statistics, but subjects often calculated or tried to calculate a mean. For example, on PreSurvey Q5a (comparing the test results for the Yellow class and Brown class), here are some comments focusing on average:

JB: Yellow class did better overall. I saw that they have more scores in the median or midrange.
SP: As a class the Brown class did better because their average score was higher @ 5 compared to the Yellow class @ about 4.5
BP: Both classes have 9 students. The Yellow class had a total of 45 and an average score of 5. The mode and median also 5. Brown class: Total = 45, Average = 5, Mode = 5, Median = 5. The two classes did equally well.

When JB referred to the "midrange", he did not mean a range of numbers but a value that is (Range)/2 , and in the context of the question is the same as the mean, median, and mode. JB was saying that the Yellow class has a higher frequency of scores at the mode. SP did not calculate Yellow's mean correctly, while BP correctly stated the mean, median and mode for both classes.

Another set of examples for this theme comes from Data & Graphs PostSurvey Q1bii, which asked for which city subjects thought was rainier:

CS: Columbus: The mean and median are higher than Portland.
RB: Columbus could be rainier because both the average and the median are higher than Portland's.

For this question, boxplots and bar charts were provided to show the sets of data for the two cities, and summary statistics were also provided. From the boxplots, subjects could see that Columbus' median was higher than Portland's, but to compare the means the subjects referred to the box of summary statistics.

As a final demonstration of measures of center captured the attention of some subjects, consider PostInterview Q8. The question showed weights for 35 different muffins from the same bakery, and asked what subjects thought their own (36^{th}) muffin might weigh. The set of data for the 35 muffin weights were shown in a boxplot (median =113.5 grams) and in a histogram (mode = 113.0 grams), and the mean was given as 113.79 grams. All the subjects below expected their muffins to weigh between 113 and 114 grams:

DS: Because... here [On the boxplot]... your median is right at, like, 113 and a half, so … And here [On the histogram] your mode is at 113
GP: Well, the median [He points to summary statistics] is 113.5 grams.
RL: Well, I'm looking at the mean, [He points to summary statistics] and I'm looking at the mode, in this case, which really stands out...
JM: Well, I look at the median as it's written out [In the summary box], and I also look at the amount of muffins at 113 [On the histogram], which is… you know, the mode is actually 113 too, and here [On the boxplot] it's also the median …

JM seems to have misread the median on the boxplot as 113 instead of 113.5, but the main point in the two responses above is that they show a focus on the average in comparing graphs.

ii) Focus on Range or Extreme Values - PreSurvey Q6 invited a comparison of student heights at two different schools, and the question elicited many responses that focused on range. For the two bar charts shown in Q6, School A had a wider range of student heights than School B. However, in School B, the heights of the adjacent bars varied up and down more frequently than the smooth rise and fall of the bars in School A. Some subjects discussed ranges without using numerical descriptions, making it difficult at times to tell if the subjects were referring to the range of bar heights or to the range of student heights:

JX: Because there is a wider range of difference in the heights recorded in [School] B than [School] A.
LW: School A shows a broader range of heights.
MA: The range of heights from shortest to tallest is greater in School A than School B.
MG: Because they have more students of different heights or a greater range of heights

Other students explicitly used a numerical description in mentioning the range of student heights:

JB: [School] A has more variability in height of students, ranging from 145-165. School B ranges from 148 to 162
MM: School A has a broader range. Because the heights vary from 145-165 in Graph A whereas in School B only 148-162.
SW: The question was which graph shows more variability. If you look at School A, the heights range from 145 to 165 with only 147 not included. School B has a range from 148 t0 162 with 161 missing. School B does not vary as much in height.

In each of the three responses just given, the subjects made a connection between variation and the range shown in the distributions of data. The connection they made is that more variation is synonymous with a wider range. I found that this connection was typical for many subjects. Often in an interview setting when I asked subjects what they meant by "more variation", one of their first reactions was to point to the range.

On PostInterview Q5, subjects had to compare graphs between Class A (50 samples of size 10 from the Small Jar) and Class B (50 samples of size 100 from the Large Jar), and some of their responses showed an attention to the ends of the ranges:

DS: Because they have fewer on the ends [She points to the ends on Class A]
SP: These lower numbers are a little surprising, for both of them.
EM: In my opinion, you might get a few more 9s, and maybe a 10.

GP: Well, I think it's pretty hard to get these 2 and 9s… I mean, really hard. See, the thing is, this [Class A: Small Jar] is a wider range, it seems like, than this [Class B: Large Jar]

DS and SP focused on the lower extremes, while the EM attended to the upper extremes. GP talked about both lower and upper extremes for Class A.

iii) Focus on Shape – The key to this theme is that responses needed to include descriptors of the shape of the distribution, and I allowed for both verbal and nonverbal communication of these descriptors. Nonverbal communication included the written responses on the surveys, of course, but also included gestures made during the interviews. By gesture, I mean that subjects were using their hands to convey an idea (such as the shape or spread of a graph) that was not always accompanied by words. Drawing a horizontal line in the air with one's hand, for example, could be a physical way of communicating a uniform distribution. There were many examples of subjects using gesture in the way I have described, to tell me how they saw or wanted the data distributed. More typical were the written or stated descriptors of how a graph looked skewed or symmetrical, or should look like a bell.

In PreSurvey Q3, subjects predicted results for 50 samples of size 10 from the Small Jar and then in Q4 they graphed the predicted results. Some subjects explicitly mentioned a shape for their graphs:

GP: Top of the pyramid is 6 or the most probable and it just cascades down.
MA: I see the result as a bell curve, since there is greater chance of getting more red than yellow, but getting ALL red is not likely either.
MG: Because in random sampling, shouldn't they fall into a bell curve?
RL: A bell curve represents the most likely scenario – the extremes aren't seen often, the average is seen the most often.

In class, I also heard other students echo GP's use of "pyramid" to refer to an inverted-

"V" shape, which also connoted the idea of a bell curve to many.

In fact, as a gesture, holding one's hands to illustrate an inverted-"V" was one way that subjects tried to signal the shape of a graph. Consider the PreInterview Q5, which showed the same set of data graphed in three different line plots, with the difference being in the scaling along the horizontal axis. GP's comment below is about Graph 1, and the other comments are about Graph 3:

GP: [Graph 1] This one...you just see that 75, you see, kind of…Kinda sloping up to 75 [Motions with his hands, makes inverted "V"]
JM: [Graph 3] It's tight. [Shows hands coming together] And it has a nice look to it [Shows hands in an inverted "V" shape>.
RL: [Graph 3] The graph 3 looks like a pretty good bell curve. It's even symmetrical! It's great.
DS: [Graph 3] Because it has a bell curve.

In GP's case, his hand motions (what I am calling gesture here) go along with his sense of how the data rises on either side of the modal value of 75. With JM, although he never referred to Graph 3 as a bell curve, he said it had a "nice look." His gestures seemed to indicate that what is "nice" to JM is the symmetry around a central peak. Other gestures that I saw for this question included the waving of hands to signify the way data fluctuated up and down across the graph.

Sometimes just certain elements of the shape stood out to subjects, such as the heights of the tallest columns in a bar chart or histogram. For example, on the Data & Graphs PostSurvey Q1, the graph for Portland's normal monthly rainfall has tall bars for the winter months and shorter bars for the summer months (denoting heavier and lighter rainfall). Several students took note of the shape for the Portland data by emphasizing the dominant winter months:

SL: Adam was noting the taller bars in Jan, Nov, Dec...They are the tallest bars on the whole chart
DS: He is looking at the "peaks" of the bar graph to come to his conclusion
SC: He was probably looking at which city had the tallest bar...Actually, Portland has the top THREE highest amounts of rain in December, January, and November
LT: By just looking at the graph it seems that Oregon has very high peaks of rain.
SP: Portland has the 3 highest bars of the graph, making it look as if Portland has more rain

The main reason I've included the above responses in the theme focusing on shape is because of the descriptive language, such as "taller" , "highest", and "peaks". Language showing how subjects attended to visual features of the graph convinced me that the shape of a distribution was a key theme in comparing and evaluating graphs.

iv) Focus on Spread – Originally I conflated shape and spread in the same theme, because I think the two characteristics of the distribution often go together. Spread has to do with the way that data clusters close to a center, or is spread out from the mean, or is concentrated at various intervals within the range. I separated the themes of shape and spread because I noticed, particularly with questions having to do with boxplots, that some responses clearly focused more on the spread of the data and less on the visual aspects (or shape) of the distribution.

Of course, some subjects actually used the term "spread" in their responses, as the following comments from PreInterview Q5 show:

DS: Well, Graph 2 you can see that... it's kind of spread out, and it's not as, at a glance it's not as, like, you kind of your eyes go "wooaahh" [She dramatically waves hand from one side of table to the other]
EM: Graph 2, I don't know. Graph 2 to me is too spread out, so I'm... I like seeing the Xs next to each other, so I can compare them easier, whereas Graph 2 is kind of spread out and I can't really read.

SP: I guess this one [Graph 3]... seems less spread out. And so , it'd be like "Oh, look how close the graph is" This Graph 2 shows it more spread out, and it'd be like "Wow! 68 all the way over here, to … 95" [Hands moving across the range]

The three subjects above didn't offer much additional detail to suggest just what they meant by the term "spread", although DS and SP's gestures imply that the range might be influential to their thinking about spread. When the PreInterview had been administered, we hadn't yet introduced in class some different ways of talking about spread.

The rainfall graphs in the Data & Graphs PostSurvey elicited many responses about spread. The following examples relate to Q1ai, when subjects were asked to write about possible causes for weather differences. In addition to talking about causes (which will be discussed as a part of the aspect of *Interpreting* variation), subjects also noted differences in spread of rainfall:

SC: So no matter which way the wind blows, Ohio gets SOME rain – It's more evenly distributed over the entire year

RL: Portland gets quite a bit of rain in the months it gets ANY rain, and very little in the summer. Columbus gets a more steady, predictable pattern of rainfall (less variation).

CM: Portland gets most of its rain from October through May, and very little from June through September. Columbus gets most of its rain March through September, but gets at least 2" per month for the rest of the year.

When CM wrote about when Columbus "gets most of its rain", she is used a naive form of spread since she doesn't mention how much is "most". Later in the research, subjects made references to percentages, such as the upper 75% of the data, to help quantify where they saw data clustered.

The Interquartile Range (IQR) on boxplots applies the middle 50% of the data,

and is one measure of spread. In Data & Graphs PostSurvey Q1bii, subjects referred to the IQR in discussing which city was rainier, Columbus or Portland:

DM: Portland, because the IQ range is higher and we tend towards massive rains in the winter and much less in the summer. Columbus is more steady
LW: I believe Portland to be rainier, because the inner quartile range is greater.
SZ: Portland, because of the interquartile range being more

The IQR was still a fairly new concept for most subjects at the time they completed the Data & Graphs PostSurvey, and it wasn't clear to me if the three subjects above were just thinking that a larger IQR translated into a rainier city. I've listed the three subjects' responses as examples to show how references to spread occurred in student reasoning.

More clarity in use of boxplots to comment on spread came in later in the research. In discussing predictions for the 36th muffin on the basis of the data set shown in PostInterview Q8, DS said

DS: Um, well, this one [Boxplot] you can see more… I think it, real clearly that 50 % are really clustered between 112 and a half and 115 and a half, and so, you go "Oh, most of 'em…you know, the middle 50% , have a very small range of weight"

Her comment exemplified the theme of spread, as she made reference to data being "clustered" and also to the "small range" for the middle half of the data. On the same question, EM and SP used both the histogram and the boxplot to comment on spread:

EM: Ummm, I'm gonna expect my muffin to weigh… I'm gonna go with the boxplot answer, of somewhere in the 50% - middle 50% range – I'm gonna expect it – and, plus, looking over here at the histogram, and that it does seem within, like, 112 to 115.5, it seems like that seems to be a concentration of data… I'm going to think that it's probably going to be in the interquartile range, of – um, like, 112.5 and 115.5 [Using the boxplot]

SP: Well… How much would I expect my muffin to weigh? Well, I'm guessing that it could be anywhere in between, somewhere around where the bulk of this data is, [Circling a central part of the histogram and the boxplot> probably … So I would expect it to be somewhere between, like, 112.5 or something to 115.5 [Corresponding to IQR]

Note how EM talked about the "concentration" of the data while SP used the term "bulk". Both terms suggest to me a relative grouping, and both appeal to the theme of spread.

C] Making Conclusions about Graphs: From the questions profiled so far in this aspect of *Displaying* variation, it is clear that opportunities were given to subjects to come to some sort of decision in the face of data presented graphically. Which class did better, what city is rainier, and which graph shows the data better – All are examples of situations that invite a conclusion. So, too, are questions that I asked in the Interviews about whose class had real or fabricated data. However, I wasn't as interested in the actual conclusions that subjects made as much I was interested in the reasons they gave and the emphases they made which had to do with an understanding of variation. For this dimension of making conclusions, I paid attention to three themes – How subjects emphasized making decisions in context, how they emphasized the consistency or reliability of the phenomena depicted by the data, and how they emphasized the level of detail or usefulness of different graphs. I'll explain these three themes next.

i) Emphasizing Decisions in Context - The key idea behind this theme is that, in considering graphs, subjects volunteered comments about the context of the data, suggesting that context was an important consideration in making a conclusion. For example, in the PreSurvey Q6 when deciding whether School A or School B showed more variation in student heights, both JL and RL's comments related to the context of student heights:

> JL: It does not indicate the gender of the students. Girls tend to be shorter than boys and there may be more girls at School A.
> RL: School A may be more homogeneous with regard to ethnicity, which is a big factor in determining height.

While the above responses also suggest *causes* of variation, I have listed JL and RL's comments here because they give a good example of emphasizing decisions in a *context*.

The context of PreInterview Q5 was a repeated-measurements experiment designed to test car brakes. In making conclusions about which of three graphs the subjects thought best displayed the data, RL's comment suggests that graphs can be used for different purposes:

> RL: Oh, I'd go with [Graph] 3, because the [Graph] 2, is too, it's spaced out...It's hard to pull it together, it's hard to say something about it, and people generally make graphs so that they can justify what they have to say.

The idea he gets across is that the context for which the graph would be used has something to do with what the user wants to say. For the same question, DS and SP shared a concern over the context of brake testing:

DS: Well, I think I would say, I would go with graph 1, because it's a little more specific on the inches, which could be a life-saving difference. And that having your, quite a few tests come up 82 or above could mean that they'd want to go re-adjust brakes. Graph 3 would be good if you were just kind of doing averages, but, I think that with brake testing that you need something more detailed.

SP: I dunno, with the ranges, it just seems like the range is pretty large [In Graph 3] And so… What is it? 70 to 79, especially when you're talking about braking distance. It seems important that you know more individually as opposed to clumping them together in 9 inches [Intervals]

I liked DS and SP's comments because they attend to the importance of having good brakes, which creates a context link to the next theme emphasizing consistency and reliability.

ii) Emphasizing Consistency or Reliability – In the process of making conclusions about graphs, many subjects referred to the consistency or reliability of phenomena. For example, they wrote about the consistent rainfall in Columbus or how car should brake consistently. In the MAX wait-time scenario for PreInterview Q8, RL remarked: "Looks like the Eastbound is more reliable." With some of the responses, it seemed that consistency also may have been a term used in reference to the shape of the distribution. However, in the examples I've selected, the focus is on how subjects make declarative statements about consistency, as if they are concluding something about the phenomena under consideration.

I'll start with examples from the car brakes situation on PreInterview Q5. In the interview script, I introduced the term "consistent" as a part of a subquestion: "If the engineer wanted to suggest that the car was fairly consistent in its braking power,

which graph would you suggest she use, and why?" Admittedly, this phrasing plants the word in the subjects' minds, and therefore is no surprise that they repeated the term back to me. I felt justified in using the term mainly because past experience has shown me that, for adult college learners, "consistent" is a common term that made sense to most. Also, I needed subjects to think in terms of the goal of the engineer for the purpose of the question. The instructive component for my research was not just that subjects talked or wrote about consistency or reliability, but in the way that they reasoned along this theme as a part of making conclusions. Here is what some subjects said about the car brakes:

DS: And so, mostly, it does consistently brake between 70 and 89 inches.
JM: We look at something like this [Graph 2], it looks much more inconsistent.
GP: Probably Graph 3, to show that it's more consistent…But, you know, if she showed them Graph 2, it would look like the car was really not being very consistent in its braking
EM: Umm. Let's see. I think that Graph 3 actually tells me ... I get a better sense of where that car is generally braking, or where it's consistently braking. So I can see that , you know, four times between 70 and 79, and four times between 80 and 89, and so... I get a sense of that , where it's usually braking.

In the last response, EM used three descriptors for braking, and they are (in order): generally, consistently, and usually. She reasoned from the histogram, with 8 data points in the middle two bins (70-79 and 80-89) out of a total of 12 data points. The central two-thirds of the data being within 70 to 89 inches told her about what the usual braking distance was.

In the rainfall comparisons of Data & Graphs PostSurvey Q1, I found many instances of the theme for consistency or reliability. The examples I've chosen came from different subquestions of Q1, but the responses all convey a similar idea:

RF: [Q1aii] Ohio is more consistent during the all year.
EM: [Q1bii] Columbus is more consistently rainy by looking at the boxplot. It's interquartile is smaller and reflects less change.
RF: [Q1bi] I think also that in Columbus it rains more throughout the year because the graph shows that the number are more consistent it is why looks more compact, pretty much is almost the same rain all year.

EM and RF's comments also reflected the theme focusing on spread in evaluating and comparing graphs, as evidenced by EM's reference to the interquartile range and RF's descriptor of the data as "compact". The main sense that I get from the above three responses is that a conclusion is being made, and the conclusion is that Columbus is a consistently rainy place (at least in comparison to Portland). On the basis of the graphs, such a conclusion is reasonable. Subjects had other terms and phrases to suggest this theme:

JM: [Q1ai] Columbus has rainfall evenly dispersed throughout the year.
MM: [Q1bi] It seems like Columbus has a constant concentration of rain
BP: [Q1bii] I personally think Columbus is rainier because the rainfall is more constant.

It seemed clear to me that an emphasis on consistency or reliability was a theme that came through in subject responses as they made conclusions about graphs.

iii) Emphasizing Level of Detail or Usefulness: When I had subjects *producing graphs*, one of the themes within that dimension had to do with technical details: the type of graph used and also the attention to scales along the axes. When students were evaluating comparing graphs, some of their responses emphasized the levels of detail offered by different graph types. Also, different graph types seemed more helpful to different students. For example, in the rainfall comparisons of Data & Graphs PostSurvey Q1, some students were more influenced by the bar graphs, and some by

the boxplots. In illustrating this theme emphasizing level of detail or usefulness, I'll first be using responses to PreInterview Q5 (about the car brakes' data shown in three different graphs). Then I'll share some responses from PostInterview Q8 (about the data for 35 muffin weights shown in both a boxplot and a histogram).

For the PreInterview Q5, Graph 1 was a line plot which only contained actual data points along the horizontal axis scale. As RL commented, "Graph 1 is very factual. It reports only the [actual] values and it takes the literal value very seriously." Since the axis was unevenly spaced, it was not surprise to me that several of the subjects did not find Graph 1 very helpful:

> EM: And then, Graph 1 also could tell me that, except that, since it puts an X for each particular number, I don't... The impact of where it's braking is lessened for me
>
> JM: Well, it [Graph 1] goes from 68 to 70, then 70 to 75, and 75 to 80, and then 80 to 82… I don't like that one, that's a little confusing for me
>
> SP: Well, the first graph, which doesn't a ton of sense to me, but, she just wrote just the distances that she got. She didn't write anything in-between, and so you're just getting like, 68 jumping to 70, to 75… And so it's just sort of, doesn't represent what would be in-between those

The responses above do a good job of illustrating this theme, since they all attended to the detail (or lack thereof) in Graph 1, and the usefulness is also addressed. DS, however, liked Graph 1:

> DS: At a glance, it's [Graph 1] easier to see, if something's presented in a concise, efficient manner, you can look at it and go, okay, most of the times, it broke, you know, 68 to 75, but it did have these trials that were higher [She shows extremes with her hands]. And it's, you know, there aren't a lot of extra numbers in there [Graph 1], which is good, you just have the numbers that it broke. That the brakes worked.

I found DS' response very interesting in that she gave a very clear reason *why* she found Graph 1 useful. She felt that Graph 1 presented the data without "a lot of extra

numbers".

It didn't seem that by the time of the PreInterview DS felt other graphs might do a better or worse job of presenting the variation in the data set. Graph 2, for instance, was a line plot similar to Graph 1 except that Graph 2 had an evenly-spaced axis. Most subjects found Graph 2 helpful in terms of the detail offered:

> JM: So when we look at Graph 2, it goes inch by inch. So it really gives you a good…Well, it shows us exactly where each [trial] landed…where it actually happened, you know…
>
> GP: Graph 2 seems to be more, shows visually better, than the others...just showing the variations that are in the distances that she, while she was braking. You almost see, like, the distance...
>
> RL: But it [Graph 1] doesn't really show the relationship as well as Graph 2, which says, okay, we're going to make a very even graph, and , so that when you look at it you get much more of a sense of what were the facts on the ground. And so it's [Graph 2] a more visual, it's more intuitive visually, more useful visually, graph.

While the above three subjects clearly express the usefulness of Graph 2, DS had the opposite opinion:

> DS: Where this other one [Graph 2], that has too much going on, and you go ["Huh?" = She gives a confused look]

Graph 2 showed more information about where the data points fell in relationship to one another, but such a relationship was either confusing or not relevant to DS. For her, Graph 2 held too much detail. However, the detail is important to give a visual sense of the variation. For instance, when JM said that Graph 2 showed "where each [trial] landed" , the same could be said of Graph 1. Graph 1 also showed each data point. What JM really meant is given away by the next part of his response, that Graph 2 showed what "actually happened". And what "actually happened" is not just that data fell at certain places (as in Graph 1), but that the data was scattered along the

axis (as in Graph 2).

Graph 3 was a sort of histogram with stacked "X"s instead of contiguous bars. JM and EM found Graph 3 useful, although JM qualified his endorsement:

> JM: Graph 3, of course, groups up in ten-inch segments, and groups them up like that. Which is okay, but depending on how critical you have to measure something, maybe ten inches is too much, if you're measuring in inches.
> EM: Ok. Well, when I can see where the distances fell, [She points to Graph 3] and if they're closer together, then it's easier for me to see how they compare, I guess you would say

GP commented on the grouping in Graph 3, and suggested that it could be used to trick the reader:

> GP: With Graph 3, you don't really get the feeling of that much of a distance between the numbers. Graph 3 really seems compact. Well, they have the groups, they have 'em grouped together, um, from 60 to 69, groups like that… I think it [Graph 3] would fool them more…

I thought GP gave a very clear indication of how Graph 3 obscures detail, and RL expanded on the same idea:

> RL: Graph 3...uses such broad grouping categories...and so it suggests a broader range has been included when, depending on your take, it could also be considered a misrepresentation.
> I: Does this misrepresent the data?
> RL: It doesn't misrepresent the data, but it does suggest more flexibility in interpretation, I guess.

RL's comments gets at the very point of this theme, which is that different graphs impart different levels of information and are useful for different purposes. In considering the general purpose of graphs, he noted that "maybe their fundamental purpose, if not a major purpose, is to visually express something usefully that does not take a lot of brainpower to derive."

While responses PreInterview Q5 gave a good illustration of the meaning of this theme, I also want to share responses on PostInterview Q8 in this section. One reason is because these responses further illustrate the theme emphasizing level of detail and usefulness, but they do so with a boxplot and a histogram. Another reason is that, while PreInterview Q5 was based on a question used in previous research (Watson et. al., 2000), PostInterview Q8 was a new contribution of mine. I'm not aware of any other studies that have gathered data on EPSTs in comparing boxplots and histograms. JM and RL were quick to note the visual power of the histogram, and how the mode of 113 grams attracted their attention:

JM: When you look at the histogram, right away, you know, it pops out: Boom, 113. The histogram is really easy, graphic display for just about anyone to see, it's 113 is the one that shows up quite often.

RL: This 113 mode is very salient [He points to histogram], it really leaps out, whereas it's not represented as such on the boxplot…

RL rightly notes that mode is not visually present in the boxplot for Q8. Subjects commented on how, in general, frequencies are not a component of boxplots:

SP: This [Boxplot] is just showing where the center half of the data is, and then, where it begins, where it ends…So you're not really getting any levels [Frequencies?] of how much is there, you're just getting that there WAS one there…

RL: Well, we also see on the boxplot, what the range is, but I can't look at this [Boxplot] and find out if there were , you know, half the muffins were 110! I just know that there was a 110-gram muffin, but I don't know how many, and vice versa on the other [Graph?]. So outside of that middle 50%, there's very little that I can glean from what's going on.

JM: Well, I can see from the boxplot that the low point is 109, it doesn't tell me how many, of course, that's one thing. It just tells me the low and the high number, and 50% of them fall within this range

The common thread in the above three responses is that boxplots don't usually tell

how much data is at any given value. The histogram, on the other hand, provided frequencies:

> JM: Whereas when I look at the histogram, I, you know, I can see every muffin just about, and how much it weighed. And if I was really concerned with each muffin, I'd know from that [Histogram] really well.
> EM: It [Histogram] actually graphs out each, each time that a certain weight came up, and so I can see more variation there
> DS: Well, because it [Histogram] has each thing detailed out, so you can see how many are exactly which weight, where this [Boxplot] gives you the general range for you know, the percentage of the numbers
> SP: Well, I think this one [Histogram] shows you the greater variation… Because you're getting each individual number, along this line,

RL had a nice way of summarizing the way he saw the differences between the two graphs:

> RL: Well, when, uh… I think the boxplot requires more interpretation. It's not quite as accessible. I look at this [Histogram] and it's very easy to compare one thing next to another, whereas here [Boxplot] – What this is really giving me is a lot of information on SOME of the data. And this [Histogram] is more complete, more thorough.

It was clear that, given the opportunity, subjects had much to say about the level of detail and subsequent usefulness of different graphs. This theme, along with the themes emphasizing decisions in context and emphasizing consistency and reliability, comprised the dimension of *making conclusions about graphs*.

[3] Interpreting Variation

A] Defining Variation: This dimension addresses what the term "variation" means to subjects, and it became clear through the research that variation had a multitude of different but related meanings. For example, a review of the responses already shared in this chapter shows how subjects thought variation had to do with the

way data was clustered or spread out, similar or different. In one situation variation was associated with the range, and in another situation variation was connected to the frequencies shown in a histogram. As GP said in PostInterview Q8 about the histogram showing the muffin heights, "Here [Histogram] you see more variation, 'cause of the ups and downs of the graph." What I found in the data was that responses fell into two distinct themes. The first theme concerned definitions and descriptions, and the second theme concerned examples. To illustrate the two themes, I'll use responses from the two questions that explicitly asked for a definition and examples, and these questions were asked on the PreSurvey.

i) Definitions and Descriptions – In response to the question "What does the word 'variation' mean to you?", most subjects' response mentioned having differences or changes. The key idea was that things were not the same:

TO: A difference of one object as opposed to the next
JM: There is variety or differences.
SP: The differences between things in a group.
DS: Changes over time

The emphasis I saw in the above responses was on the simple presence of differences or changes, which fits well with the description of variation I gave in Chapter One. Some subjects also emphasized differences in connection with making choices, and they stressed having different options or alternatives. Other responses emphasized the degree of difference or change:

CM: Degree to which something is different
JL: The degree by which a number can change, less or more
SC: It's more like all the different things that can be slightly or greatly different from what you are studying.

A last group of responses connected variation to math, similar to JL's response above:

AL: I'm not sure, it has something to do with equations
BP: How far something deviates from the average
LW: The difference or distance from the norm
RL: A measure of how a given piece of data compares with the average of similar data.

The last three responses show previous experience with statistics. I asked the question about the meaning of variation on the PreSurvey but did not ask the same question at the end of the research, and now I wish I had. However, it was clear through their responses that a broad web of meaning for the term variation occurred throughout the research. Data were described as being clustered or scattered, concentrated or widely distributed. Graphs were described as compact or spread out.

I had expected my subjects to have at least an everyday, common definition of the term "vary" and its linguistic forms (variation, variability, etc.). In the process of predicting possible outcomes for different scenarios, they frequently used another common term, "random". Randomness, as pointed out in Chapter Two, is linked to variation in the fundamental sense that appreciation of one concept should accompany an appreciation of the other. Therefore, I also asked on the PreSurvey what the term "random" meant to them, and one of the dominant characteristics I saw in responses was that randomness implied a lack of pattern:

BP: Sporatic, having no pattern
DS: Not patterned
JX: With no pattern
LW: Without a given or set pattern.

Another characteristics I saw in the description of randomness was the effect of

unpredictability. That is, random events were seen as unpredictable. LW said it well: "One cannot predict what will come next by previous experience."

ii) Examples – In the PreSurvey, when I asked subjects to "give an example of something that varies", the main characteristic of the examples I got in response had to do with the weather or other natural phenomena:

JM: The weather. The shapes of rocks. Snowflakes.
CS: The temperature varies every day.
DM: Sunshine in Oregon in January
GP: The weather changes it's look.
MA: The amount of daylight we experience throughout the year
JB: Temperature in the Spring
RL: Sea level.

Another characteristic concerned people or personal characteristics:

SP: Weight, height, hair color of a group of people.
MG: The height of students in a class.
JM: People's attitudes.
MM: My mood sometimes. My music taste.

The survey and interview questions all reflected many different examples of variation, and in the context of data and graphs I had examples such as weather, muffin weights, train wait-times, and car stopping distances.

Subjects' responses to the related PreSurvey task "Give example of something that happens in a "random" way" suggested contexts of sampling and probability. For instance, SP wrote: "If I put a quarter in a gumball machine, the gumball I get is random." Her example related to the candy sampling tasks asked in the surveys and interviews. Other examples of random events suggested by subjects included:

SL: Powerball, maybe. Roll of dice, flipping a coin
MG: Pulling names out of a hat

JL: Selecting a bouncing ping pong ball with a number listed on it from a hot-air lottery spinner (for lack of a better word)

Through the class activities and research tasks the subjects experienced many other examples of , but even at the beginning of the research project there were some reasonable definitions and examples of variation and randomness.

B] Causes of Variation: Occasionally subjects speculated about causes of variation on their own accord, but there were also several questions in which I specifically asked for subjects' conjectures about causes. For example, during the class activities on sampling, I asked why they thought results were not all the same. Two themes that I delineated for responses about causes are naturally occurring causes and physically deliberate causes, which I'll explain next.

i) Naturally Occurring Causes – This theme includes randomness as a reason for variation in sampling and probability situations. It also includes the reasons that subjects gave for weather differences between Columbus and Portland. Every subject had at least one possible reason, and many subjects' responses listed multiple reasons, such as:

SX: Geography of the two cities cause their different rainfall patterns. Portland probably gets the higher rainfall in winter months because weather systems from the Pacific get caught between the Cascades and the Coastal Range. Maybe Columbus is too cold in the winter for large quantities of rain. In the summer it rains more in Columbus because moisture comes from the Great Lakes (I think?)

JM: Columbus gets more rainfall in the summer months. Thunderstorms and low pressure accounts for this difference. Portland's winters are mild and wet, Columbus' colder temperatures account for more winter snow. The Pacific ocean has a very large effect on Portland's climate.

I was impressed at how well thought out some of the responses like JM and SX's

were, and overall it seemed that writing about causes of variation came easily to the class as a whole in this context.

Similarly, in the traffic death rate question on the Data & Graphs PostSurvey, I included reasons such as different speed limits, drivers' age requirements, or road conditions in this theme because those are normal, routine reasons which might account for variation in the data. Here are some sample responses for causes of variation in the traffic death rates between the South and the Northeast regions of America:

AL: Weather, speed limits, types of roads, age of drivers
DM: Older roads in the South , older vehicles, weather, more cars on the road, as opposed to busses & trains.
JM: The legal age to operate a motor vehicle could be lower in the South, contributing to the higher death rate.
LW: Rural roads versus urban areas. The age of drivers. The years of experiences driving. Road conditions. Drivers education requirements.

LW's comment about rural roads versus urban areas was echoed by some other subjects, who expanded on the difference:

JL: More rural areas in the South with highways and faster speeds than the NE. The faster speeds, the more likely the accident will result in death. The hospitals may have better critical care facilities.
EM: In the South there are probably longer stretches of highway, more space between destinations, more people falling asleep or not paying attention on long drives – higher speeds
BP: Maybe because there are more flat, long, open roads in the South – People have to drive farther to reach places and there is opportunity to drive faster. The NorthEast has more Metropolis and things may be closer, more freeway driving, not as much country road driving?
DS: I think people in Northern states, although rates are based on same # miles, drive actual SHORTER distances each time they drive because population would be more condensed. That would decrease chances of death. Because less chance of accident on short drive than long drive.

Again, I was quite impressed at the depth of thought given to the above responses. A number of subjects attributed the cause of variation to alcohol consumption, and I considered this as a cultural factor (or at least the subjects' impression of a southern culture):

DS: Maybe the people in the Southern states drive drunk more.
EM: Maybe even a correlation between education level and drunk driving?
GP: Higher incidence of drinking and driving.
JM: Perhaps a higher incidence of drinking under the influence of alcohol in the South.
LW: More alcohol consumed due to the heat causing more accidents.

There were even more responses than those above listing the South in connection with alcohol consumption, and I was somewhat surprised at what seemed to be a bias coming through in student responses. SR plainly said: "ALCOHOL! From my own personal experience and biases, Southern people as a whole drink and drink more, and are more careless also, but that is my own experience."

In addition to the causes for variation in weather patterns and traffic rate deaths, many reasons for different MAX wait-times – reasons such as the precision of the watches, or how the middle schoolers may not have had their watches perfectly synchronized – seem a natural part of the process of data gathering. Similarly, on the repeated muffin-weighing question on the PostInterview, having crumbs fall off a muffin as it gets weighed seems a normal occurrence.

ii) *Physically Induced Causes* - However, having crumbs fall off a muffin is different from taking a bite out of the muffin to deliberately introduce variation. Someone actually suggested the "bite" cause for the muffin repeated-measurement

scenario, an example of a physically deliberate cause. The main characteristic of this theme is that someone or something acts in a purposeful way which introduces variation into a situation as a result of the act.

Several physically deliberate causes were volunteered by subjects in both sampling and probability contexts. In sampling, several students addressed the nature of the candy mixing. While it may seem that "mixing" is an inherent and natural part of the sampling scenario, some responses emphasized the way that the hand might grab the candies, making the situation seem as if the person doing the sampling was causing the variation. In PreSurvey Q1b, GP wrote that he thought a person would get different results when drawing samples from the small jar, "because you will probably grab differently and the candies are shifting to different places." One of GP's emphases is on the person doing the drawing. It seems from what GP wrote that if one did not grab differently, and tried to grab candies the same way each time, closer results would occur. GP's response is a good example of stressing physical causes in a sampling environment. Other responses emphasized the physical environment, such as the way "yellow candies could be bunched together in the jar", or "how many red or yellow happen to be in the area you grab". The way the candies were mixed and subsequently chosen seemed to be concern to some subjects in the sampling situations.

In the probability contexts with the spinner, there was a strong perception from some students that a person doing the spinning can cause more or less variation by virtue of controlling or influencing the spinner. In fact, the spinner attracted more comments about physically deliberate causes than the other random devices such as the coin or the die. Here are some spinner comments from PreSurvey Q8, which

asked if there was a 50% chance of winning a game involving two fair spinners (each spinner was half-black and half-white):

MM: Only if the spinner starts spinning in between both is it a 50-50 [Chance]
RF: I think a lot depends on how you spin
SW: I think it depends somewhat on where the spinner is started from

SW above also had an idea about flipping a fair coin 50 times in PreSurvey Q7a. She thought that perhaps the coin would land heads-up "a little more than half 'cause it started on heads", although she qualified her response by saying "I have no idea really." Other examples of physical causes will be brought out in the case study discussions, because some students changed their emphases on physical causes from before to after the class interventions, and I'll be sharing some of these comparisons later in this chapter.

C] Effects of Variation – In this dimension, my focus was on the effects of variation *on the subjects*. Variability was inherent in the tasks used in this research, and subjects had different levels of understanding of what constituted reasonable expectations in the face of this variation. I do not suggest that subjects themselves are necessarily aware of the effects variation has upon their responses, but I hypothesize the following two effects: The first is the effect on subjects' perceptions, and the second is the effect on subjects' decisions.

i) Effects on Perceptions – Variation inherent in situations can affect how subjects perceive those situations. Two characteristics of responses within this theme suggested what students "know" or "perceive". First, many students said they knew that reality was different from theory. Second, when considering results from a

variable situation, many subjects said they knew that results could be anything.

In perceiving that reality is different from theory (the first characteristic of the *effects on perception)*, subjects mostly commented about probability theory. For example, when considering a single sample of size ten from the Small Jar on PreSurvey Q1a, MG wrote: "If you take a random sampling of any population, you should get a proportional representation." MG therefore has a good sense of what probability suggests, and later in Q1c ("Six Trials"), MG put all sixes for his choices. But in Q2a ("Range 6"), he put "3 to 8" for his range, and later explained that "if they are being selected randomly, there shouldn't be the same number coming out each time." It seems as though the reality of the situation, at some point, comes into focus for MG and contrasts with the expectations based on probability. Responses from other students on PreSurvey Q1 include:

DS: [Q1b] Because probably outcomes aren't for sure outcomes
RL: [Q1b] Reality does not obey the estimates of probability
SR: [Q1c] You are dealing with chance, like gambling. In theory there is probably an answer…a 6-4 chance each candy picked is red. But if you do it for real, 100 times, the numbers change but the ratios do not.

The key idea in the above responses was how probability says one thing, but what really happens is another. While the above responses were from the sampling context, there were similar responses in the probability context. Here are two examples related to the coin-flipping scenario of PreSurvey Q7:

DM: [Q7b] In all likelihood it would probably be different, but statistics say again it should be 25
RL: [Q7c] While 25 flips are likely to be heads, in reality some variation is likely, so my numbers represent a range that averages 25

Notice how both DM and RL approached the same theme from opposite directions. DM has the theoretical side covered by what "statistics say," and on the reality side she notes the likelihood of differences. RL talks about the likeliness of the theoretical (25 heads), and on the other side is the reality of variation. Another term some subjects used for the theoretical expectation was the "perfect" result. For example, in considering the expected value of 25 blacks for a sample of 50 spins of the fair spinner, DS said that "it'll still be rare to get the perfect 50%". Similarly, SA knew that she wouldn't get the expected value every time in drawing samples from the Small Jar, saying: "That would be too perfect." The idea seems to be that in a perfect world, results would match the theoretical prediction. In the real world, variation happens.

The second characteristic for this theme was reflected by comments about how results could be anything. I saw this second characteristic as an extension of the first, because if subjects perceive that the expected value won't always occur, then sometimes they reasoned that any of the other outcomes in the sample space could occur. Logically, an event is in the sample space can in fact occur, but the responses displaying this second characteristic seem to ignore relative likelihoods of events. Consider SP, who predicted the following six results for six samples of 50 flips each of a fair coin in PreSurvey Q1c: 2,3,10,16,22,25. Her predictions are low, and not on both sides of the expected value of 25 heads. Moreover, her lowest results of 2 and 3 heads are extremely unlikely. Her reason for her choices was that she "just chose randomly – anything is possible." In the Sampling PostSurvey environment of drawing samples from the Large Jar, there were other responses similar to SP's:

SZ: [Q2c] It is random selection, anything can happen.
SR: [Q3c] Even after the month of lessons on stat. & probability, I still feel that it is luck and fate of what each turn will pull...anything is possible.
JM: [Q4b] Anything is possible.

Granted, the subject perceptions described within this theme - about reality versus theory, and how results could be anything – relate just as well to randomness and uncertainty as variation. Indeed, many of the responses for this theme echo traits of intuitive probabilistic thinking reported in earlier studies, most notably the Outcome Approach. Semantics do come into play when talking about uncertainty, randomness, and variation. I linked subjects' perceptions to variation, given the broad definition of variation introduced in Chapter One. There will be differences (variation) in results, and subjects therefore perceive that reality is not the same as what theory predicts, in fact results could be anything.

ii) Effects on Decisions – While the previous theme focused on what subjects "knew" or "perceived", this theme concerns the subjects' decisions or ability to make decisions. Some subjects claimed it was difficult to know what results would occur, and that they couldn't predict or decide. Other subjects expressed a lack of confidence in making inferences.

An "I don't know" type of answer was often given by subjects who were asked what might happen in sampling and probability situations. Sometimes subjects also used the "I don't know" line of reasoning when explaining their answers. Guessing was also listed in response to many different questions, such as the following examples from the PreSurvey:

SP: [Q1c] I just made a guess – even though there is no way to systematically prove my guesses
SL: [Q2c] Total guess. I have no background to predict from
CS: [Q3b] Guess. Have no idea
AL: [Q3b] I would be totally guessing if I wrote #'s down. I don't know how I would figure this out.
SL: [Q7b] Couldn't hazard a guess, or could but it would be random

An idea behind the "I don't know" and "I'm guessing" types of responses is not that subjects are not *able* to guess or predict, but that they cannot *know ahead of time* if their predictions are correct. The following two responses from the PreSurvey directly address the difficulty in making predictions:

AL: [Q1b] You can make a prediction, but not a concrete answer as to what color you will pick.
LT: [Q1b] Always getting six red candies is hard to predict.

One of the uses of statistical reasoning is to make inferences. For some subjects, making inferences was difficult, and it seemed that the variation inherent in situations led to subjects' claim of difficulty in predicting. When LT writes (as above) that it is "hard to predict", it seems that what she is really saying is that it is hard to predict and then have the prediction match with the actual outcome.

In the interview setting, many subjects also clearly showed their difficulty in decision making, with some questions eliciting long, protracted attempts to reconcile the reality of variation with the theory of prediction. The two themes for this dimension are connected in the sense that how a subject perceives a situation influences the ease and confidence they have in making predictions or decisions based on that situation.

D] Influencing Expectation and Variation: Two themes I saw in connection with this dimension were quantities in sampling, and the number of samples taken. I'll illustrate the two themes and also how subjects seemed to relate the themes to influencing expectation and variation.

i) Quantities in Sampling – This theme arose from questions about drawing samples from the Small Jar and from the Large Jar. The key idea was how subjects' responses (or parts of their responses) focused on the numbers of the candies in the sample or in the jar rather than emphasizing the ratio. Other researchers have used the term "Additive Reasoning" to describe the focus on the sheer size or numbers used in sampling situations (Shaughnessy et. al., 2004). Here are some examples of responses from PreSurvey Q1, concerning samples from the Small Jar (60 Red/ 40 Yellow):

MM: [Q1a] Because there are more red candies than yellow
MA: [Q1a] I stand a greater chance of pulling more red than yellow, because there are more of them to begin with in the [jar]
RF: [Q1b] Because if I had more red I have more probability to get more of these
SW: [Q1c] The odds are that each classmate would have more red because there are 20 more reds to begin with.

The common theme in the above responses is that there are more reds than yellows in the jar. In the absence of any additional information, their responses above beg the question of whether the likelihood of getting a red candy has more to do with the numbers in the jar or with the proportion. That is, the responses show more that the subjects are influenced by *quantities* (the numbers of candies) than they are influenced by *proportion*. For instance, SW explicitly mentions how "there are 20 more reds", showing an additive strategy.

In the next set of examples, LT also uses an additive strategy, claiming there's "only 20 more reds". These responses come from Sampling PostSurvey Q1, when subjects were reasoning how many samples from the Small Jar they would expect to make in order to get 0 or 1 red candy in their sample:

LT: There is not much of a difference between 60 Red and 40 Yellow. There are only 20 more reds than yellow.
SA: I'm sure it has something to do with the fact there is so many more reds than yellow
MG: The likelihood of getting only yellow is low because there are so many more red than yellow.
SL: The red has higher chance 'cause there are more.

I placed this theme concerning the number of candies in the dimension of *influencing expectation and variation* because I wondered if the likelihoods were seen by subjects as influenced by the quantities and not relative quantities. For instance, since SL suggests above that red has a higher probability of being chosen because there are more red candies to begin with, perhaps she would therefore reason that the probability would increase if the numbers increase (but the ratio stays the same).

On the Sampling PostSurvey Q4, when justifying the predictions for 50 samples taken from the Large Jar, again there were further suggestions of additive reasoning:

LW: Since there are more red than yellow I believe it more likely for the trend to push higher rather than lower.
MG: Because there are so many more red than yellow, they will be more likely to pull more than 60 rather than less
SA: You have a better chance of pulling all reds than pulling no reds because there are 200 more reds than yellow in the jar.

Thus, the sheer size of the samples and populations featured prominently in subjects' reasoning about both Small and Large Jars. There are some useful ways in which

moving from a Small to Large Jar does influence the distribution, but I hypothesize that many subjects only think in terms of greater numbers leading to higher or lower likelihood. For instance, they may think that trying to draw a red marble from a 60 Red/ 40 Yellow mix is an easier (likelier) task than trying to draw a red marble from a 600 Red/ 400 Yellow mix.

ii) Number of Samples – There were four characteristics I saw within this theme. The first characteristic was that more samples has no effect on the probability associated with individual results. In stressing the stability of the underlying proportion, communicating how the ratio doesn't change no matter how many trials are performed, RB expressed the argument this way: "No matter how many people take a handful, the odds will always be the same because each handful is replaced before the next person draws." In PreSurvey Q7 when subjects considered samples of 50 flips of a fair coin, some wrote as follows:

JL: No matter how many times he flips, the odds are the same
SP: No matter how many times he flips it, the chance is still ½
AL: I don't see how the chances of getting heads will change if he does more sets of 50 flips
SR: Still he has a 50/50 chance on each flip and on each group of flips

While the above observations true, the responses were often used to justify an expectation of no variation in results from repeated samples. In other words, there an assumption that because the number of samples does not change the underlying ratio, whatever was result expected for one sample should be extended to all samples.

The second characteristic was that more samples would yield more variation, and in this sense variation was used as a synonym for range. In other words, doing more samples would extend the range in both directions. On PreSurvey Q2, I asked a question that invited a comparison of ranges for a smaller and larger number of samples. Responses included the following:

CM: The range will increase with increasing attempts
JL: The more people that do the experiment, the more varied the results.
RL: As the number of trials goes up, so expands the range of possible outcomes towards the extremes.
MG: I think (?) there should be a larger range of variation (from the mean) as the number of samples increases.

Another way that students had of expressing the view of an expanding range was to say that more samples gave more chances to get extreme values, or as JX put it on Probability PostSurvey Q2: "The more sets done, the more likely you will get less likely results."

The third characteristic was that more samples also gave more chances to actually attain the expected value, and related to this characteristic was the notion that the average of a set of trials should be (or be closer to) the expected value. Often the principle of the Law of Large Numbers was implicit in responses, and in one instance the Law was explicitly stated. The following examples are in response to Probability PostSurvey Q1b, as subjects consider more than one sample of 50 spins of the fair spinner:

GP: The more times, the closer it will be toward 25
LW: The more times he spins, the closer he will actually get to the 50/50 chance
SA: The more he spins the closer the results will match the probability (1/2)

SX: Because there are more spins, the variation will be less than with only 50 spins, hence closer to ½ the # spun (50)
MG: It will be even closer to 25 because of the Law of Large Numbers

In SX's response, when she says that "the variation will be less" with more spins, she may be using the term variation to refer to the relative clustering of results around the mean and not to the absolute range.

The fourth characteristic was that more samples give a better picture of the underlying distribution. For instance, distributions become more normal, and subranges – such as the range capturing the central 90% of the results – shrink relative to the absolute range. In the Probability PostSurvey, here are two examples of this characteristic concerning the distribution and an increasing number of samples:

JL: [Q1b] I think the sample results would get tighter, the grouping would accumulate around [the expected value]
RL: [Q2] The more trials run, the more normal the distribution, but the chance of outliers also increases

Thus, taking more samples was thought to have no effect on the underlying proportion, and to increase the chances of expanding the range by attaining extreme values. Also, more samples improved the chances of actually attaining the expected value, and more results would cluster around the expected value, affecting the shape of the distribution.

Summary

This section has shown what I mean by each of the themes that make up each dimension for each aspect within my evolving framework. The framework addresses my first research question by providing a comprehensive structure to characterize EPST's conceptions about variation. It must be reiterated that the framework allows

for responses to fall across more than one aspect, dimension, or theme, depending on the complexity of the response. Researchers using this framework will occasionally come across responses that only exemplify a single theme, and will frequently encounter multi-thematic or multi-dimensional responses. Most of the examples of student responses in this section have been excerpts of longer responses, for the purpose of highlighting the meaning of the themes. The framework was informed by the entire corpus of data on all instruments, although I deliberately chose exemplifying responses more from the Surveys and less from the Interviews. In the next section, I apply the framework to compare six individuals' conceptions of variation from before to after the class interventions, and I focus on their responses to the Pre and PostInterviews.

Individual Cases

To answer my second research question, I used the evolving framework as a lens to view the thinking of six subjects who each participated in two interviews. I looked for significant ways in which subjects' conceptions changed or remained the same as the subjects progressed through the research. The framework helped characterize my findings, and the case studies are organized according to the main aspects of *expecting, displaying,* and *interpreting* variation. I'll describe the main ways that each of my six cases showed stability or shifts in thinking within each aspect.

Of the eleven subjects interviewed, there were three females (SP, EM, and DS) as well as three males (GP, JM, and RL) who were selected to be the six case studies for this research. They were selected mainly because their collective responses

spanned all the themes of the framework. Moreover, they had no problem sharing their thoughts in the interviews, and their narrative provided vivid illustrations of their thinking. Each of the six cases participated in two videotaped interviews, with each interview lasting about 45 minutes. The PreInterview was given within two weeks of administering the PreSurvey, and was conducted before formal instruction on probability and statistics in MET 2 had begun (recall that the first four weeks of the quarter in MET 2 was spent on geometry). After four weeks of doing lessons and activities on probability and statistics (this time frame corresponded with weeks 5 – 8 in the ten week quarter), the PostInterviews were conducted. To illustrate the comparisons for each case, I'll mainly use responses to a subset of questions from the Pre and PostInterviews.

The interview questions were summarized in Tables 4 and 8 of Chapter Three, and they are found in their entirety in Appendix B. I chose a subset of the interview questions for case analyses for three reasons. First, the cases' collective responses on these questions spanned all the themes of the framework. Second, the questions themselves spanned the three contexts of sampling, data and graphs, and probability situations. Third, the questions were specifically constructed to so that PreInterview questions were isomorphic to PostInterview questions. By isomorphic, I mean that the questions were phrased similarly or addressed similar ideas. I've reorganized Tables 4 and 8 to show how questions matched up, and also to assign nicknames that will help identify the questions used in this section (see Table 12).

Table 12. *Isomorphism of Interview Questions*

Pre	NickName	Scenario Involved	Post	NickName	Scenario Involved
Q1a Q1b Q1c	One Sample Several Samples Six Samples	Small Jar : 60R/40Y (Samples of 10)	Q1a Q1b Q1c	One Sample Several Samples Six Samples	Large Jar : 600R/400Y (Samples of 100)
Q2	Compare Lists		Q2	Compare Lists	
Q3	Graph: 30		Q3	Graph: 30	
Q4	Graph: 300		Q4	Graph: 300	
Q6	Causes: Train	21 Times Recorded (One MAX Train)	Q6	Causes: Muffin	20 Weights Recorded (One Bakery)
Q7	Compare Graphs		Q7	Compare Graphs	
Q8	MAX Wait-Times	10 Wait-Times Each (Two MAX Trains)	Q9	Muffin Weights	12 Muffins Each (Two Bakeries)
Q9 Q10 Q11	One Sample Who Cheated? Six Samples	Six-Sided Die (Samples of 60)	Q10a Q10b Q10c	One Sample Compare Samples Six Samples	1:1 Spinner (Samples of 50)
*	*	*	Q11	Compare Lists	
Q13	Likelier Graph?	2:1 Spinner (Samples of 60)	Q13	Likelier Graph?	

(* Along this row, the Post Q11 "Compare Lists" was in the Probability context and did not have a direct counterpart in the PreInterview. Post Q11 is similar in structure to Post Q2 and Pre Q2.)

In Table 12, I've created nicknames that reflect the content of the questions, and in general the Pre and Post questions can be matched by their nicknames. For example, "One Sample" for the Small Jar on Pre Q1 is similar to "One Sample" for the Large Jar on Post Q1. "MAX Wait-Times" (Pre Q8) is similar to "Muffin Weights" (Post Q9), and "Who Cheated?" (Pre Q10) gets at the same essential idea as "Compare Samples" (Post Q10b). There is one question in Table 12 (Post Q11) which does not match directly across to a counterpart in the PreInterview. Post Q11 had subjects "Compare Lists" in a probability context, but on the PreInterview I only had subjects "Compare Lists" on Q2 in a sampling context. Despite differences in context, useful comparisons of subjects' responses were still made between Pre Q2 and Post Q11.

In next presenting the case studies, I'll use the following structure. First I'll introduce the case, summarizing upfront the main points of stability or shifts in a student's thinking. Then I'll describe further details according to each aspect of the framework.

The Case of DS

DS was a very energetic individual who readily expressed opinions and thoughts on all the questions. She had taken MET 2 the previous quarter with Steve, and had also taken a prior course in probability and statistics at another college, saying she "loved it." On the PreSurvey, for her initial definition of variation she had said that variation meant "changes over time," and cited her mood as an example of something that varies.

Summary: It was clear from the PreSurvey and PreInterview that prior to the class interventions, DS already had a good grasp of the basic ideas involved in probability and statistics. She showed a facile use of proportional reasoning, and usually expected results of repeated samples to vary. She also gave reasonable ranges for predicting results of six samples, but was generally wide on her ranges for thirty or more samples. She expected ranges to increase as the numbers of samples increased. Lastly, in attending to graphs DS referred to center, range, and shape of the distribution.

DS corrected herself at two key points during the PreInterview: Once when she first thought that all tens was a good guess for "One Sample" of the die tossing (Pre Q9), and again when she initially thought Group B had a realistic graph in "Likelier Graph?" (Pre Q13). In both instances she changed her mind on the basis of

her "Won't be Perfect" reasoning strategy. I was surprised that she misidentified "Graph:30" as actual results for Pre Q3, since I would have expected her to say that it also looked too "perfect."

The main changes I noticed in DS's collective responses were that she had more complex responses in the PostInterview. In particular, she tended to use more descriptive language in the PostInterview than in the Pre when talking about the variation in situations, such as how data was "clustered". She attended to more features of the distribution (average, range, shape, and spread) in the Post than in the Pre, often incorporating several features in a single response. DS had some fairly reasonable ideas and good ways of communicating during the PreInterview, but she added depth to her responses and expressed herself even better in all aspects during the PostInterview.

Expecting: I'll use DS's responses to PreInterview Q1 as a starting point for this discussion. In "One Sample" from the Small Jar, DS predicted a result of 6 reds:

I: [Pre Q1a] How many red do you think you're going to get?
DS: I think I'm going to get 6 red.
I: Why do you think that?
DS: Because 60% of the 100 are red, and 6 is 60% of 10. So, it's not for sure. The odds are.
I: What do you mean, "it's not for sure" ?
DS: Well, 'cause, I could get a different amount of reds, and a different amount of yellows.
I: Right now you're saying 6
DS: Six is the best shot.

Her response was very reasonable, and she reasoned proportionally yet also acknowledged the possibility of variation. In "One Sample" from the Large Jar, she

predicted a value close to the expected value of 60 reds:

I: [Post Q1a] How many do you think are red?
DS: Sixty-four
I: Alright. Why do you think that?
DS: Because, odds are, 60% are red, and you're probably not going to get exactly 60%, just because of the variability of the blind drawing… so 64 is close.

She again used proportional reasoning and considered the possibility of variation, and she used the same "odds" language as in the PreInterview. She actually expected variation from the mean, and had a prediction which she knew was "close" enough to be reasonable. What I found interesting was that she explicitly said that she probably wouldn't get 60 reds, and she gave a reason. While the expected value of 60 reds may be the most likely outcome for one sample, DS's response seemed to acknowledge that the likelihood of actually getting 60 reds is small.

For predicting "Six Samples" from the Small Jar on Pre Q1c, DS had a reasonable list of "4, 5, 6, 6, 7, 7", and I wondered if she would just multiply by ten when moving to the Large Jar. However, on Post Q1c her "Six Sample" predictions were "56, 58, 60, 61, 62, 64" , which all fall within a reasonable range for sampling from the Large Jar. In the probability context, however, DS initially expected no variation when considering the number of times each face of the die would show in "One Sample" of 60 tosses (Pre Q9):

I: [Pre Q9] What do you think is going to happen, for these faces?
DS: I'll just go, ten of each [Writes down all tens]
I: Why do you think those numbers are reasonable?
DS: Because… one out of 6 is going to roll up a "1", and one out of 6 will roll up a "2"…But then, going back to that question about picking the colored candies and the 6,6,6,6… That's kind of … 10,10,10,10 is kind of like 6,6,6,6… so it probably will vary somewhere.

I: Well, put what you think, Debbie.
DS: Ten is as good a guess as any.
I: If you rolled it 60 times, that's what you think you're going to get?
DS: Sure. [Seems pretty confident] It's as good as any guess.

DS clearly was influenced by proportional thinking, but I noticed that DS reflected back to "Six Samples" from the Small Jar. Her expectations from Pre Q1c seemed to conflict with her expectations for Pre Q9. However, an outcome of ten was "as good as any guess" for one face, hence good enough for all faces.

When I showed DS the supposed results of dice tossing on Pre Q10 ("Who Cheated?"), she was quick to identify Lee's list of "10, 10, 10, 10, 10, 10" as unbelievable:

I: Explain your reasoning, please.
DS: Well, because really, the 10, 10, 10, would be so unusual that it would come out that way.
I: Ok, it would be so unusual, and yet that's exactly what you said you thought might happen [Turns back to Q9]
DS: Well, I don't really think that it's going to happen. It's a guesstimate… It's an educated guess.

For DS, all tens was a reasonable guess and she listed all tens as an a priori expectation. When faced with the same result of all tens as a supposedly a posteriori result, she was quick to see that all tens was just not very realistic. She went back to Pre Q9 to change her "One Sample" result list from "10, 10, 10, 10, 10, 10" to "6, 8, 10, 10, 12, 14", saying all tens was too "perfect", and unlikely to happen in real life. I'll comment more on her line of reasoning against reality being "perfect" in the *interpreting* aspect. When asked to evaluate Lynn's list ("10, 11, 10, 10, 9, 10") as a part of Q10, DS said:

DS: And then Lynn, I don't think that out of 60 rolls that there's not enough variation, between what came up how many times...he [Lynn] only went over one and under one. Where really, chance could probably have a broader range.

I thought DS had some good reasoning in evaluating the different lists in the, and she clearly connected a narrower range with having less variation. When I asked DS to predict how many fives would result in each of "Six Samples" of sixty tosses (Pre Q11), she wrote "6, 8, 10, 12, 13, 14" saying she "liked those numbers". She also indicated she thought that with more samples, she would have chosen a broader range. The tendency of wider ranges in larger samples also arose for DS in the *interpreting* aspect.

DS never again listed a string of all identical numbers when given the opportunity to predict results. For instance, in the probability context on the PostInterview, she listed a reasonable "21, 23, 25, 26, 28, 29" for "Six Samples" of the spinner (Post Q10c). When explaining her choices, she included some proportional reasoning, and then said:

DS: So I have one 25 here. And then I have a few scattered close to 25, but not 25...'Cause there's gonna be variation, because the spinner CAN land anywhere, but probably on average it'll be close to 25.

DS included many themes in her above response. She appealed to the notion of distribution by describing how results are "scattered close to" the expected value, and she acknowledged that individual results and the average of a set of results will vary.

Displaying: DS suggested that the thirty supposed results of "Graph: 30" on Pre Q3 were actual results:

I: [Pre Q3] Which of the following do you think is most likely?
DS: Oh, I think… those could have been the [actual] results
I: Why do you think that's the most likely?

DS: Because six is our odds-on favorite, and they just didn't have a lot of variation when they picked out their candies.
I: What makes you say "they didn't have very much variation" ?
DS: Because here's six, and they're only one away from six, on each side [She pointing with her finger on the graph, tracing out the range]

Thus, DS was comfortable with the unlikely narrow range portrayed by "Graph: 30" (see Figure 22), but she didn't have too much to say in terms of her justification. She focused briefly on the mode of 6.

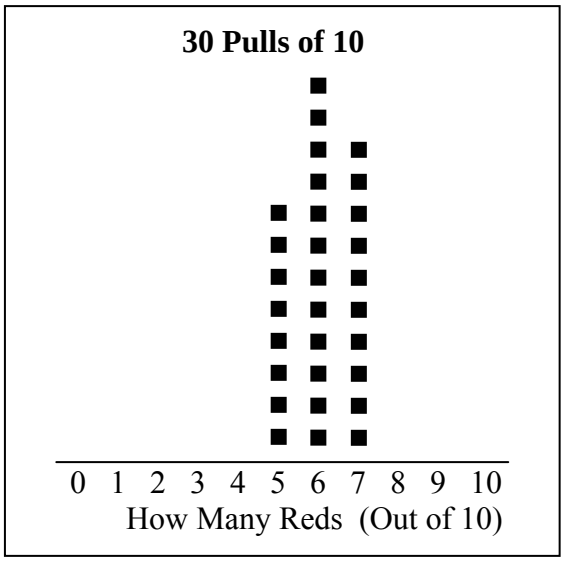


Figure 22 – PreInterview Q3 "Graph: 30"

She also attended to the range in making her evaluation, noting how the thirty results ranged from 5 red to 7 red candies. It is possible for thirty actual results from the Small Jar to look like the graph shown in "Graph: 30", but not very likely.

In the PostInterview, the "Graph: 30" for the Large Jar on Q3 really did represent actual data (see Figure 23), and DS made a correct identification of the graph as being authentic.

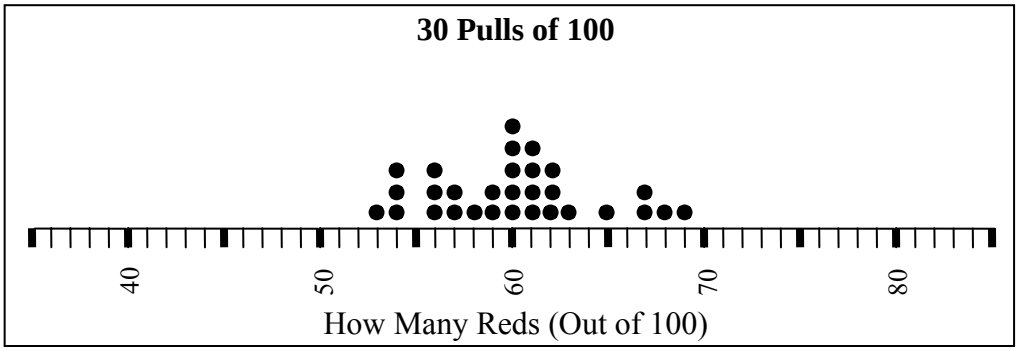


Figure 23 – PostInterview Q3 "Graph: 30"

More importantly, she offered better reasoning on the Post than she did on the Pre:

I: [Post Q3] So what do you think is most likely... that Craig's class just made up those results, or those are the actual results…
DS: No, I think those could be the actual results. [She is quick to respond]
I: Could you explain why you think [that's] most likely?
DS Well, because our, you're most common number is 60, which is the average number of reds, and then, there's kind of a cluster around that number. And then there's just a few on the edges…
I: So you like that
DS: Yeah, and then there's just, you know, a little straggle here and a straggle there [She marks the min and max]

DS's evaluation of the graph included a focus on the mode of 60. When she said that 60 is the "average number of reds" she means that 60 is the expected value, not the mean of the data set shown in "Graph: 30". She also appealed to the spread of the data by talking about the "cluster" of results around 60. Finally, she attended to the extreme values by marking them on the graph. Thus, her multi-thematic response on Post Q3 involved three themes focusing on average, range, and spread.

When DS worked on the "Compared Graphs" task in Pre Q7 (see Figure 24), she suggested that both graphs told a similar story about the duration of the train trip, "that it was somewhere between 58:30 and 59:30".

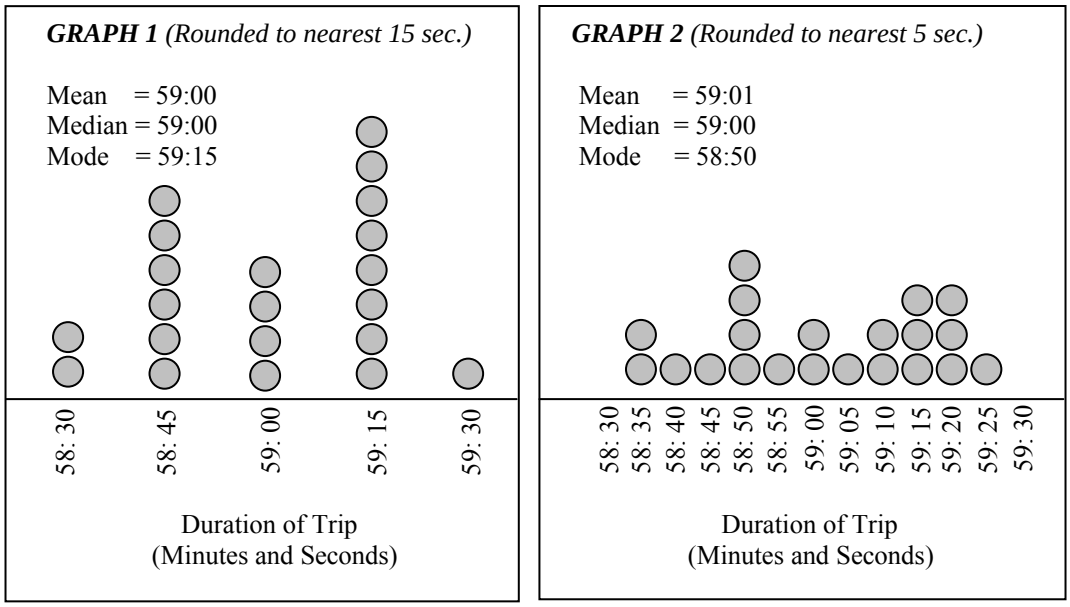


Figure 24 – PreInterview Q7 "Compare Graphs"

Thus, her initial focus was on the range. For Graph 1, she pointed to the subrange represented by the two middle bars and said that "more people experienced that time frame", and she also counted individual data points on both graphs to help her compare. What I found most interesting about DS's response in Pre Q7 was her conclusion about the two graphs that "I think I like them both. Either graph is fine." Her response on the similar "Compare Graphs" on Post Q7 involved more distributional reasoning and also a firmer decision in favor of Graph 1 (see Figure 25). She thought the two graphs on the Post Q7 told her different stories, with Graph 2 appearing more spread out:

> DS: Yeah, 'cause this [Graph 2] is kind of...it's like spread out in teeny increments, and kind of detailed like that. And also, in this Graph [2] there's so many numbers that you kinda go "Too Much!"...Like too much flatness, for it to really make a statement about how much it weighs. Where here [Graph 1] it makes more of a statement, like "Oh, probably weighs 109 or 111 – Somewhere in there."

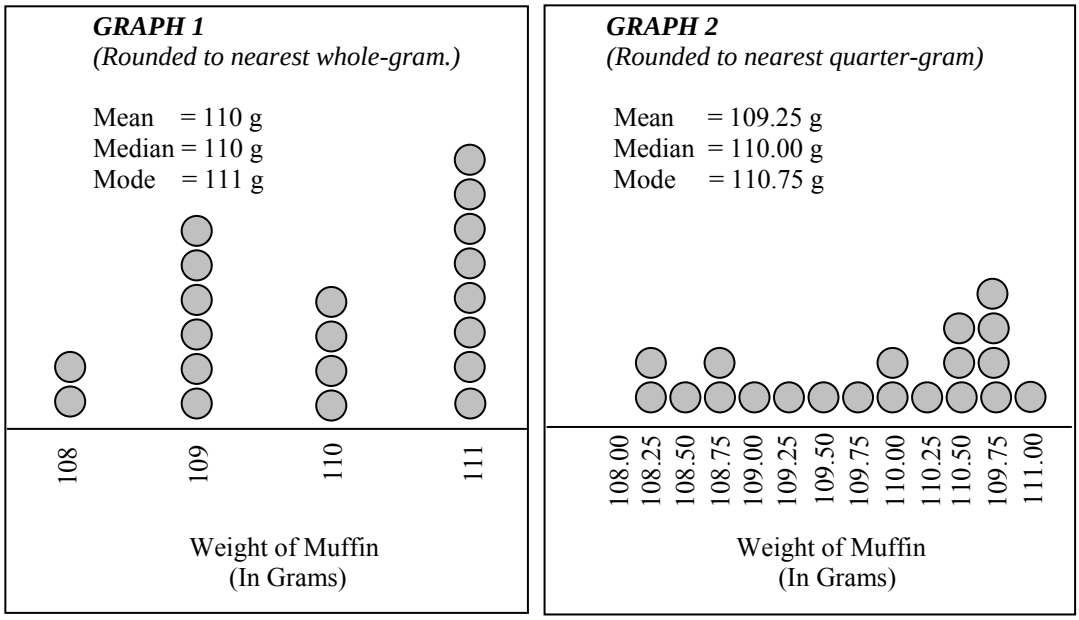


Figure 25 – PostInterview Q7 "Compare Graphs"

DS has two dimensions reflected in her response. For evaluating and comparing graphs, she attended to the shape of Graph 2, noting its "flatness". For making conclusions about graphs, she emphasized the level of detail and subsequent usefulness of Graph 1 versus Graph 2. Specifically, even though the rounding is finer in Graph 2, DS liked Graph 1 because it better conveys to her where most of the data fell.

Another example for DS's reasoning about *displays* of variation comes from the "Likelier Graph?" questions on the Pre and Post. On Pre Q13, Group B's graph is fake (see Figure 26), but DS initially said "I think Group B looks more like what I would expect." When asked why, she appealed to the shape, saying "it's that famous curve" (the graph was roughly bell-shaped).

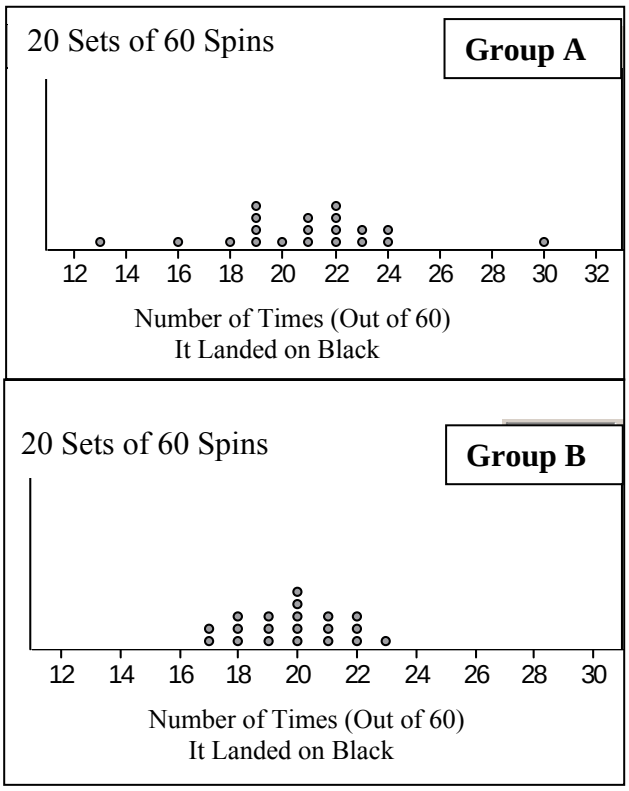


Figure 26 – PreInterview Q13 "Likelier Graph?"

She also focused on the mode, which was at the expected value of 20 blacks for a sample of 60 spins of the 1:2 (Black:White) spinner. Her comment was:

> DS: And so, most of those, in Group B, fell in that one-out-of-three...20 [blacks]. And then, you just vary a little on each side, 'cause you only have 20 sets [of 60 spins per "set"]. You don't have a lot of sets.

Finally, she focused on the range for Group B, which she liked. DS used all elements of the distribution – average, range, spread, and shape – in her evaluation of Group B, and she was convincing herself that the graph was reasonable. Then she changed her mind:

> DS: Back up. I think Group A looks more real.
> I: Now what are you thinking?

DS: Well, now I'm thinking that, you know, that it's not going to always end up in this perfect graph picture. So this [Pointing to Group B] would be, if you were going to fake a graph? This would be a fake graph [Laughs].

When pressed for more reasoning about her change of mind, her main rationale was the expectation for a more expanded range with 20 samples than what was pictured for Group B.

I thought that DS had some reasonable thoughts on the "Likelier Graph?" task in Pre Q13, and she expressed her ideas well, but in the PostInterview she expressed herself even better on this task (see Figure 27).

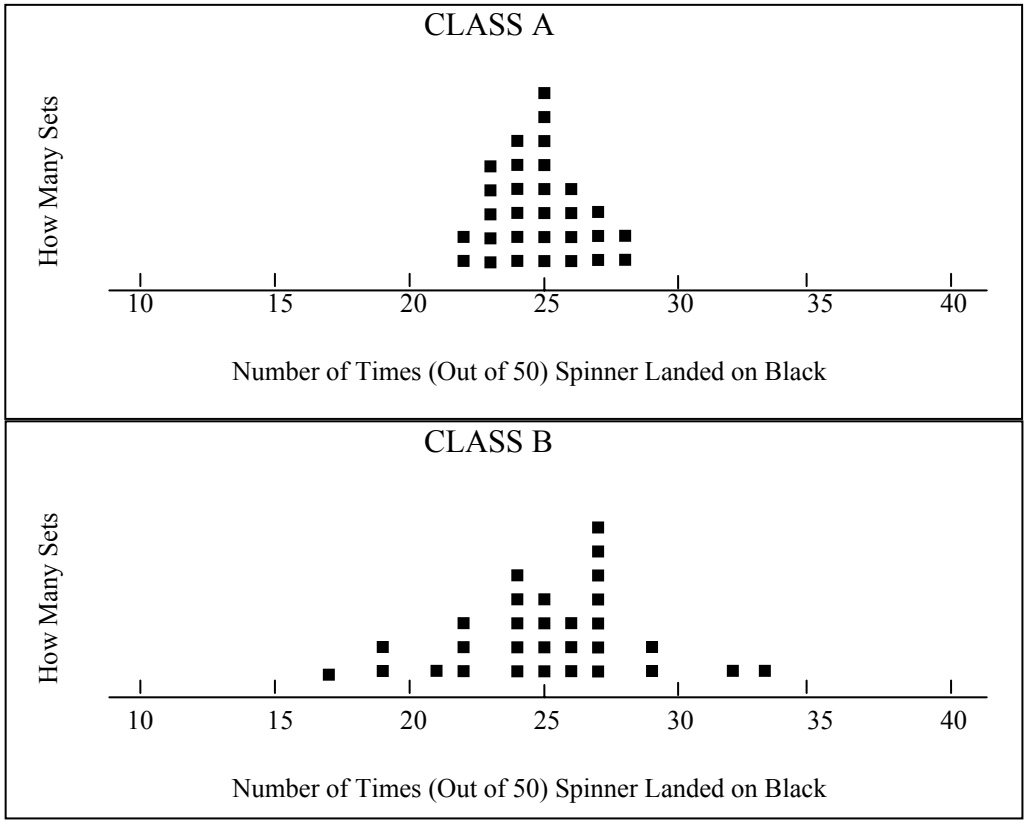


Figure 27 – PostInterview Q13 "Likelier Graph?"

This time Class A had the fake, more compact bell- shaped graph, and Class B was real:

DS: There's this...kind of, you know, bell curve [for Class A] that's kind of clustered right in the center, without much variation on either side of that 25 [The expected value]. Class B, you still have your cluster in the center...and then you have a few little odd birds [The extreme values]

DS considered shape and spread as she compared the two graphs (see Figure 27). Class A's mode was right at the expected value of 25, whereas Class B's mode was at about 27, but what DS particularly liked in Class B was the "cluster in the center". She also liked how Class B had a broader range than Class A's range of about 22 to 28. She was quick and confident to denounce Class A as a fake:

I: Do you have a sense that one class or the other is likelier to have...
DS: [Interrupts] I think Class A cheated.
I: Well, why do you think that?
DS: Well, because it's toooooo perfect. I just think somebody would have gotten, you know, under 20, or 20, or you know, 32...You know, even if it's just one person, that would be out of the cluster.

She shows her a sense of distribution by wanting a grouping close to the expected value, tapering off on either side of that value, and a reasonable sense of range.

Interpreting: DS continually referred to the "perfect" results in her reasoning on both the PreInterview and PostInterview. I saw that her sense of the "perfect" world was tied to her perception of variation. As seen in her above response to Post Q13, when rejecting Class A as the "Likelier Graph?", she claims Class A looks "toooooo perfect". Class A's graph is not completely symmetrical, but what DS meant was that the general shape and tightness of range was not realistic. The same idea of the "perfect" shape comes through in her comment about Group B's graph in Pre Q13's "Likelier Graph?", when she says that one won't always have a "perfect graph picture." Group B did have the mode at the expected value of 20 blacks, and DS was explicit about how "your odds are...1 out of 3 is going to be black, in a perfect

world."

DS's reference to the "perfect world" shows how she connects perfection to the absence of variation. She seemed to perceive probabilistic theory as predicting what would happen in the "perfect world" , while results in the real world varied away from the "perfect". I asked her to explain more about her perceptions when she had decided that rolling a die sixty times and getting each 10 of each face was unrealistic:

> DS: [Pre Q10: "Who Cheated?"] Because it's TOO perfect [Lee's choice of all 10s]. Life doesn't happen that way. It could but it doesn't. [Laughs]
> I: Why doesn't it?
> DS: Because it's… a random thing.
> I: Could you tell me what you mean by that?
> DS: That there's chance involved, so, whenever there's chance, then things won't necessarily turn out perfectly. Like, in a perfect world situation, where the dice was loaded.

I think DS's reasoning serves her well. If it were a "perfect world situation," she seems to be saying, then less variation would mean more consistently correct predictions. Chance leads to variation, both of which are related to uncertainty. The "perfect" result was clearly one number, she explains further:

> DS: It is an idea that I hold. So, 'cause I think that, um, that there IS the chance that it'll come up perfect, but there's … "perfect" is one [Holds up hand to signify one number] , and there're more things that are imperfect, like, not perfect. [Waves hands to show distribution of other numbers on either side of the "perfect" number] So, there's a lot more options for the imperfect.

Her explanation above shows why DS said for "One Sample" at the Large Jar in Post Q1a that her result probably wouldn't be exactly 60 reds. A sample result 60 reds is just one of many other possibilities. Her sense of the "perfect" went beyond just the expected value to include shape and spread, as exemplified by her responses to the

"Likelier Graph?" questions. She also commented on a list of supposed results for six samples of the fair spinner in "Compare Lists", Post Q11. About list (v), she said:

> DS: Choice (v) is good, the only thing is that it's so…Perfect...You know, 24, 24, 25, 25, 26, 26…There's not a lot of variation there, which I think there might be a little more.

List (v) is so "perfect" because it only varies by one on either side of the expected value of 25, and because it shows a uniform distribution of two samples for each of the outcomes of 24, 25, and 26.

Another example of DS's thinking in the *interpreting* aspect is how she built upon her notion of what effect taking increasing numbers of samples would have. She already thought in the PreInterview that more samples meant a widening range and she held onto that notion in the PostInterview:

> DS: [Pre Q1b: "Several Samples" of the Small Jar] The more I choose candies, the more chance there will be that I'll get different than six reds. Either fewer or more.
> DS: [Pre Q11: "Six Samples" of the Die] If we had more sets of 60, then I would make my numbers go lower than…[Showing with hands a greater range]
> DS: [Post Q13: "Likelier Graph?"] The more spins you do, I think there's more chance that you'll get… A number that varies from your, you know, further from your 25.

DS had already expressed how more samples meant more chances for the "imperfect", what happened in the PostInterview was that she added to her notion of more samples meaning a wider range. In the Post (but not in the Pre), she added the idea that more samples gave more chances to actually attain the expected value. For example, on Post Q10c, "Six Samples" of the spinner, DS said: "But the more times you spin it, the more chance that you'll get 25." Now compare what DS said above in "Several

Samples" of the Small Jar on Pre Q1b to what she said about "Several Samples" of the Large Jar on Post Q1b:

> DS: [Post Q1b] And the more times you pull, you'll have variations on each end, which might get wider, but you'll have more in the center, around the 60 number.

Whereas in Pre Q1b she mention getting "fewer or more" with more samples, in Post Q1b she includes the language of "variations" to describe a widening of the range, and she also added the distributional idea that more samples meant more near the center, "around" the expected value of 60 reds.

Finally, DS also related more samples to the shape of a distribution in the Post, but not in the Pre. A good example comes from "Graph:300" (see Figure 28), and for the Small Jar on Pre Q4 she mainly attended to the range:

> DS: [Pre Q4] Well, this ["Graph:300"] has a broader range of picks, of number picks [Her finger traces out the range on the horizontal axis]

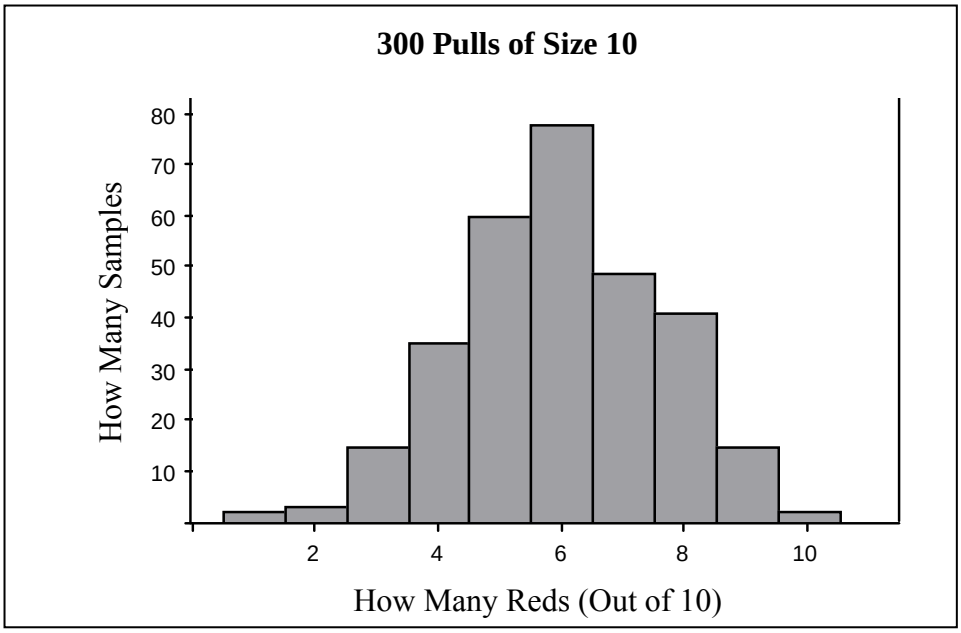


Figure 28 – PreInterview Q4 "Graph: 300"

Her answer was much more thorough on Post Q4 in evaluating the "Graph:300" (see Figure 29):

DS: [Post Q4] Well, the most common is right around 60, and then there's fewer on the edges as get further away from 60. And, in class when we did, on the computer, the more pulls you do, the more evenly shaped your graph is going to be. Where fewer pulls, you're going to have a little more unevenness in your curve

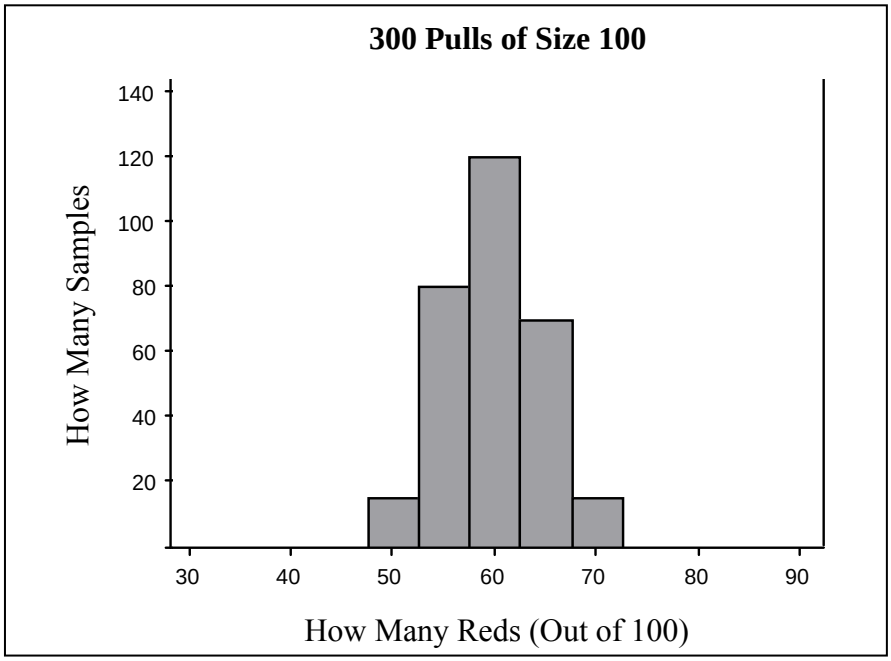


Figure 29 – PostInterview Q4 "Graph: 300"

DS's thoughts are extremely well-said, and she includes many themes such as attention to average, range, and she includes class experience in her reasoning. The focus for the *interpreting* aspects is the effect of the number of samples: More pulls means a more "even" shape, less pulls means more "unevennesss" in shape.

In conclusion, there was one instance on the PreInterview where DS had predicted all tens for each side of the die as the outcome of 60 tosses on Pre Q9 (One Sample). During the PreInterview, she changed her mind about the reasonableness of

her prediction, and never again predicted identical results. She erroneously supported fake graphs as being realistic on the PreInterview, but her later evaluations of such real-versus-fake graphs were more reasonable on the PostInterview. DS had stability in her language and reasoning from Pre to Post about how results wouldn't always be "perfect", and added to her thinking about the effect of doing increasing numbers of samples in the PostInterview. The biggest difference for DS in all aspects from before to after the interventions was a qualitative broadening in her reasoning skills. She had some reasonable ideas in the PreInterview, but her responses were deeper, more complex, and better expressed by the time of the PostInterview.

The Case of GP

GP was an effusive character who used gesture quite frequently in his explanations during both interviews. Like DS, GP had also taken Math 211 the previous quarter with Steve, but unlike DS, he had taken no prior classes in probability or statistics that he could recall. When asked how he felt in anticipation of learning the topic, he said "I'm open to it, but not really excited." His initial sense of what variation meant was "a different look to a subject," and when asked to give an example of something that varies, he wrote "the weather changes its look."

Summary: In *expecting* variation, on the PreSurvey and PreInterview GP showed a fairly naive understanding of how unlikely some extreme values were. He became more sensitive to the presence of outliers by the time of the PostInterview, and incorporated more language about what was possible or likely. He consistently thought results from multiple samples should usually be different from one another,

and also mentioned experience as a reason on both Pre and PostInterviews. The sampling activities seemed to make an impression on GP, particularly the computer simulation.

When considering *displays* of variation, on the PreInterview he made many reasonable evaluations and comparisons, and he continued to do so in the PostInterview. He was quicker to make conclusions about graphs in the PreInterview, and was occasionally less decisive initially in the Post. The biggest change for GP in this aspect was that he became much more sophisticated in his communication about graphs, and on the PostInterview he reasoned more distributionally.

For *interpretations* of variation, GP consistently focused on the physical nature of the candy mixing, although he did so to a lesser extent in the PostInterview. He also consistently used forms of the word "random" to describe many facets of variation. While he knew that the number of samples was likely to extend the range of results, this idea came out stronger in the PostInterview than in the Pre. He also considered the numbers of candies in the jar to be influential for the PostInterview.

Expecting: On the PreInterview, GP gave "6 red" as his prediction for "One Sample" of the Small Jar (Q1a), and he gave a proportional reason. Thus, I knew that GP was capable of reasoning proportionally, which was a significant finding because he also frequently relied on additive reasoning (meaning that he focused on the sheer numbers of candies). He included additive reasoning (to be discussed further in the *interpreting* aspect) when he predicted "70 reds" for "One Sample" of the Large Jar on Post Q1a. His language in answering "Several Samples" was similar on both Pre and Post Q1b, when I asked him if he would get the same results each time:

GP: [Pre Q1b] Probably not. Probably not, no.
GP: [Post Q1b] No. I mean, not every time. It could happen, but...Not very likely.

Both of his responses show the theme of possibilities and likelihoods, and on the Post he often was very explicit about results being possible but unlikely. For "Six Samples", GP's range was too wide on the PreInterview (he picked 1, 3, 4, 5, 6, 10), and he said:

GP: [Pre Q1c] It just randomly came to me. I thought if you stick your hands in there randomly, you could just pick up any number, between 1 and 10

The result of 1 red is very unlikely for "Six Samples" of the Small Jar, yet GP did not comment on the relative likeliness for that lower extreme value. He also did not seem to consider zero reds as a possibility, and he uses "randomly" to describe both a cognitive process and a physical process. In the PostInterview, his range and reasoning both improved on "Six Samples" of the Large Jar, when he chose 48, 50, 58, 62, 68, 72, saying:

GP: [Post Q1c] You want to pick around 60, kind of going a little extreme... I just picked around 60, and 10 or whatever [Waves his hands to show both sides of 60]. I mean, it could go anywhere...So I just picked more likely options.

GP knew the expected value in Post Q1c was 60 reds, but he didn't expect any of his "Six Samples" to actually be 60, just "around 60". He gives very reasonable choices. When he said "and 10 or whatever", he meant plus or minus approximately 10 on either side of 60. Even though he has the same "It Could Be Anything" kind of statement he made in the PreInterview, on the Post he stressed that he was picking results that he felt were "more likely".

The "Compare Lists" questions on the Pre and PostInterviews showed more of what GP *expected* and why. For example, on Q2 in the PreInterview, GP felt list (i) was "fine", even though list (i) for six results of samples of the Small Jar has numbers that are only 6 and above ("7, 9, 7, 6, 8, 7"). Other subjects tended to notice that the entire list seemed high. List (ii) was most reasonable ("6, 7, 5,8, 5, 4") and GP liked it because it had "less radical numbers", which was the way he often referred to extreme results. List (iii) had all sixes, which GP did not like because "all in a row would be pretty unlikely". GP thought results for multiple samples should be "random", which frequently meant different or lacking a discernible pattern. Thus, he liked list (iv) – "2, 5, 4, 3, 6, 4" – because it was "kinda random, you know, not that many radicals in there". He did not comment on List (iv) being low overall. He also liked list (v) – "3, 10, 9, 2, 1, 5" - saying:

GP: There's the 1 and the 9, that's pretty, you know...[High? Rare?] But I like that.
I: Isn't that one most like the one that you put [On "Six Samples"] ?
GP: Yeah, I kinda...I did that because, I kinda wanted to be a little radical

GP did favor list (ii) overall, but it was clear from his responses that high or low results were fine with GP, and extreme values were not a concern, but he did like results to not all be the same.

On the similar Post Q2, list (i) for the Large Jar was also high ("72, 91, 74, 63, 81, 78"). GP checked off list (i) on Post Q2 as one of several lists he liked, saying the six results "just look like a bunch of random numbers, that were picked out of a jar." Later in his response to Post Q2, GP eventually commented on the result of 91 being

"pretty rare". He still never commented on how list (i) was high overall. List (ii) on Post Q2 was the most reasonable choice ("61, 73, 56, 69, 59, 48"), and it was GP's favorite because they were "just pretty random numbers...they're all different, there's no rhythm to 'em." As in the PreInterview, on Post Q2 he did not like all the repeated values of list (iii), and again he didn't comment on how list (iv) was low overall. At the end of his consideration of Pots Q2's "Compare Lists", he commented on how list (iv) had a 34, but list (ii) had "less extreme numbers."

Towards the end of the PostInterview, on "Compare Lists for six samples of the spinner (Q11), again list (i) was high ("38, 43, 36, 26, 41, 33"). This time GP was more cautious, saying: "Um, I guess it's possible. The 43 and 41 is pretty high, but... Well, it's possible, I guess." Although he didn't comment on list (i) being high overall, he did focus on the extreme value of 43 for the spinner just as he had done for the extreme value of 91 for the Large Jar. Finally, on Post Q11 GP noticed how list (iv) was low overall ("15, 19, 11, 25, 21, 23"):

GP: [Post Q11] I'd be surprised at this one too [List (iv)].
I: Why?
GP: Well, you have the 11...Yeah, these lower numbers, but...Possible. I mean...The highest one is, there's nothing over 25, so that's pretty unlikely.

For GP, the shift to the theme of possibilities and likelihoods showed a bit more hesitancy about accepting the highly improbably extremes shown in some of the lists on PostInterview Q2 and Q11.

GP mentioned experience as a reason for his expectations on both interviews. In "One Sample" of tossing the die on PreInterview Q9, GP was one of the two cases who did not put all tens for the faces of the die. He put "7, 8, 9, 11, 12, 13" because:

GP: Well, I knew it was going to be random, and so I first looked at the 60, and I said, well, these all have to add up to 60. So I divided by 6, and I said 10 each. And then I said, well, it's not going to be happening, 10 for each one, and so I just took 2 numbers and made it so they would equal 20, like 7 and 13 is twenty...8 and 12 is twenty, 9 and 11 is 20. And so I knew that would all add up to 60

GP again ties "random" to differences, he uses some part-to-whole reasoning, and also he knows that a uniform distribution is "not going to be happening." Thus, in considering "Who Cheated?" on Pre Q10, GP was quick to denounce Lee's results of all tens as unbelievable, saying "I look at it and go: Come on, Lee! There's no way that this happened!". GP also thought Lynn's results ("10, 11, 10, 10, 9, 10") "seemed too... Balanced. Too – Not as random, or something." For Pat's results ("2,15, 10, 28, 1, 4"), GP is explicit about relying on experience:

GP: Pat...That's pretty, kinda believable, but ...[Takes his time thinking] um...Gonna...Too extreme, I guess...
I: What tells you that?
GP: Well, she only hit...with the 60 times, she only got one "5" ?
I: Oh, yeah...
GP: I mean, that's...You're going to get more "5"s than that, out of 60 rolls, you know?
I: Okay
GP: I've played board games, and I'd roll dice, and you get 5 more than that, you know

After the class interventions, GP referred to experience several times on the PostSurveys and in the PostInterview. For example, when considering different arguments for how twenty samples of the spinner might look on PostInterview Q12, GP said: "And that's when I would pull out the Phantom [Fathom] software and show 'em how this works." It was also clear from his comments in class that the activities in sampling and probability, combined with the computer simulations, had made an

impression on GP. He would point out results that had been obtained experientially as justification for what he expected.

Displaying: Whereas GP had some questionable expectations in other question involving sampling and probability on the PreInterview (and even on the PostInterview), I was surprised at how reasonable many of his ideas were in evaluating and comparing graphs. It seemed to me that he was more of a visual learner, attracted to graphs in the sense that he responded with much energy. For example, in PreInterview Q3 and Q4, when he was thinking about whether "Graph: 30" and "Graph: 300" were real or fake, he was quick to judge "Graph: 30" as fake, saying "I think they cheated". He thought there should be a wider range for "Graph: 30", but did think that the mode should be at 6 and the shape should resemble a "pyramid." He used "pyramid" several times in the interviews, often accompanied by holding his hands in an inverted "V". It seemed that "pyramid" was a way of connoting a bell-shaped distribution, and he justified his thinking of "Graph: 300" as real by saying:

> GP: [Pre Q4] I think this is more like the pyramid, what I would see. This looks more legit to me [Holds hands in inverted "V"]. It seems like it spreads out...you have a few extremes out here. and then it kinds goes up, where it is more likely in the middle here [Points to mode of 6]

GP appeals to all aspects of the distribution in one response: Average, range, shape, and spread.

On the similar Q3 and Q4 on the PostInterview, GP found it difficult to decide if "Graph: 30" was real or fake:

GP: [Post Q3] It's definitely possible. I can't see how you can say that this is...You look at it and go 'No, this is fake", you know? I just see that they're all...kind of gathering in the middle around 60. Anything is possible, you can't say 'Oh no, you guys did this wrong, you cheaters!' You know, I would say these are actual results: 'Good job, guys!' You can't prove that...they cheated. You can't.

However, on Pre Q3, GP had said confidently "I think they cheated", and I suspect that his softening of graph judgments might be linked to his sense of what was possible. Commenting on "Graph: 300" in the PostInterview, GP thought it showed actual results. His response included a focus on average and shape, and he also invoked experience as a justification:

GP: [Post Q4] The majority is over the sixty, kind of tapers off...That's usually the look of a large-number grab. You get more of that look. That's my experience.

When he said "over the sixty", he meant that the data was literally piled up above the mode of sixty.

A comparison of GP's responses to Q7 on both interviews indicates a situation where he showed more decisiveness as he "Compared Graphs" in the PostInterview than he had in the Pre. Looking at Graph 1 in the PreInterview, GP focused on the mode of 59:15, saying it was "really tall" and that "your eye usually goes to the tallest one." The mode was a visual attractor for GP. He then said "I think Graph 2 is more helpful", but as he explained his thinking, he started to do something that no other case did. GP started using his pencil to re-distribute data on the different graphs, trying to figure out how they compared to one another. He talked aloud as he shuffled data around, and seemed to come to an impasse about which graph was giving him more useful information. Then he said:

GP: [Pre Q7] I think this one [Graph 1] is easier...This one [Graph 2] gets a little confusing, you know. This one [Graph 1] is if you were going to talk about it, it'd be easier to do this one [Graph 1].

On the similar "Compare Graphs" question on the PostInterview, GP had an opposite opinion. That is, Graph 1, which had coarser rounding than Graph 2, was denounced as "totally misleading." He had some interesting comments about use of the average in either graph, saying:

GP: [Post Q7] I think the median and mean thing is kind of a tricky thing to use, in just weighing this one muffin. Because you're kinda compromising the weight, kinda thing, you know? You're just saying: You know, we didn't get one answer, so let's just...get the middle between the mistakes here...

I thought GP's ideas provided a basis for thinking about averages as a way of balancing out the variation ("mistakes") in the data. He went on to talk specifically about the rounding strategies used in generating both graphs:

GP: [Post Q7] Well, this [Graph 2] is more accurate because you're taking the less, the rounding – to the lowest quality, so you get a more accurate view of what you got. This [Graph 1] is more spread out, you know, less differences here [Graph 2] between the measurements, you can see.

GP uses "spread" to describe the range of Graph 1, which is wider than the range of Graph 2.

Just as GP talked about the mean and median being "tricky" in Post Q7, he also had some difficulty reconciling the identical averages on "MAX Wait-Times" with the differences in spread shown in the data sets (see Figure 30) in PreInterview Q8.

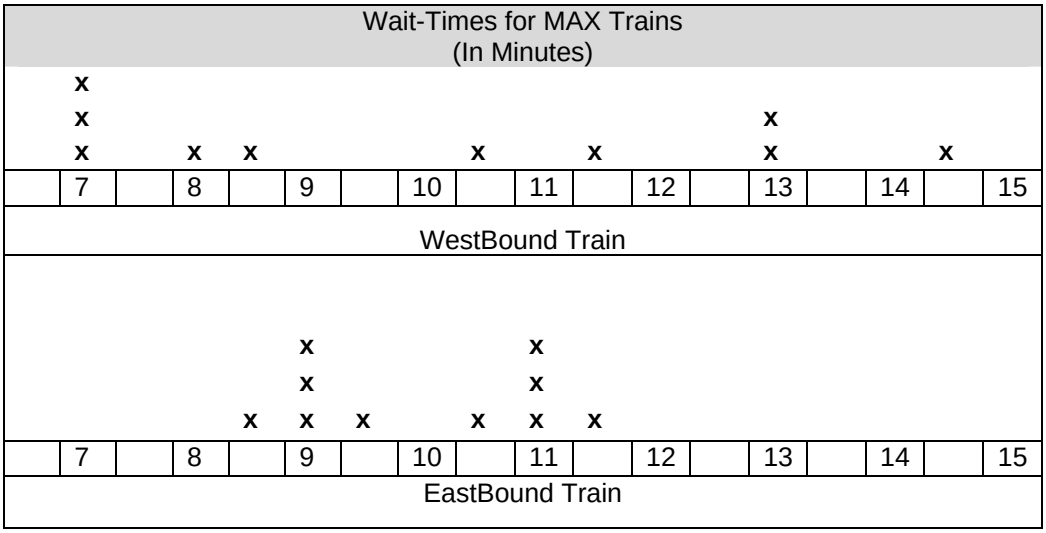


Figure 30 – PreInterview Q8 "MAX Wait-Times"

His initial impressions on PreInterview Q8 were:

> GP: [Pre Q8] Well, the Eastbound train seems to be more consistent. 'Cause their's seem to be more closer together [Pushing his hands together] This [Westbound] seems to be less consistent, since it's more spread out…[Drawing his hands apart]

His body language made me think that GP was primarily focusing on range, and when he considered the summary statistics, he seemed to experience the tension between centers and spread:

> GP: [Pre Q8] So in some ways it DOES balance out in the end, because these are balancing out as you see in these , these calculations. [Waves hands across the graphs, then points to summary stats] I'm going back and forth with these mean and median kinda things...
> I: Oh yeah? What do you mean?
> GP: Well, it's saying that they're the same in some ways, you know.
> I: Interesting
> GP: They have both the same numbers in the top and bottom [Same averages]. But then...they look different, down below [Different graphs]. So you're kinda going back and forth.

His eventual conclusion aligned with his initial impression, which he repeated, adding that the "Westbound train seems more of a...gamble."

GP correctly connected more variability with less reliability in the "MAX Wait-Times" question on the PreInterview, and he reasoned similarly on the Post Q9 about "Muffin Weights" (Figure 31).

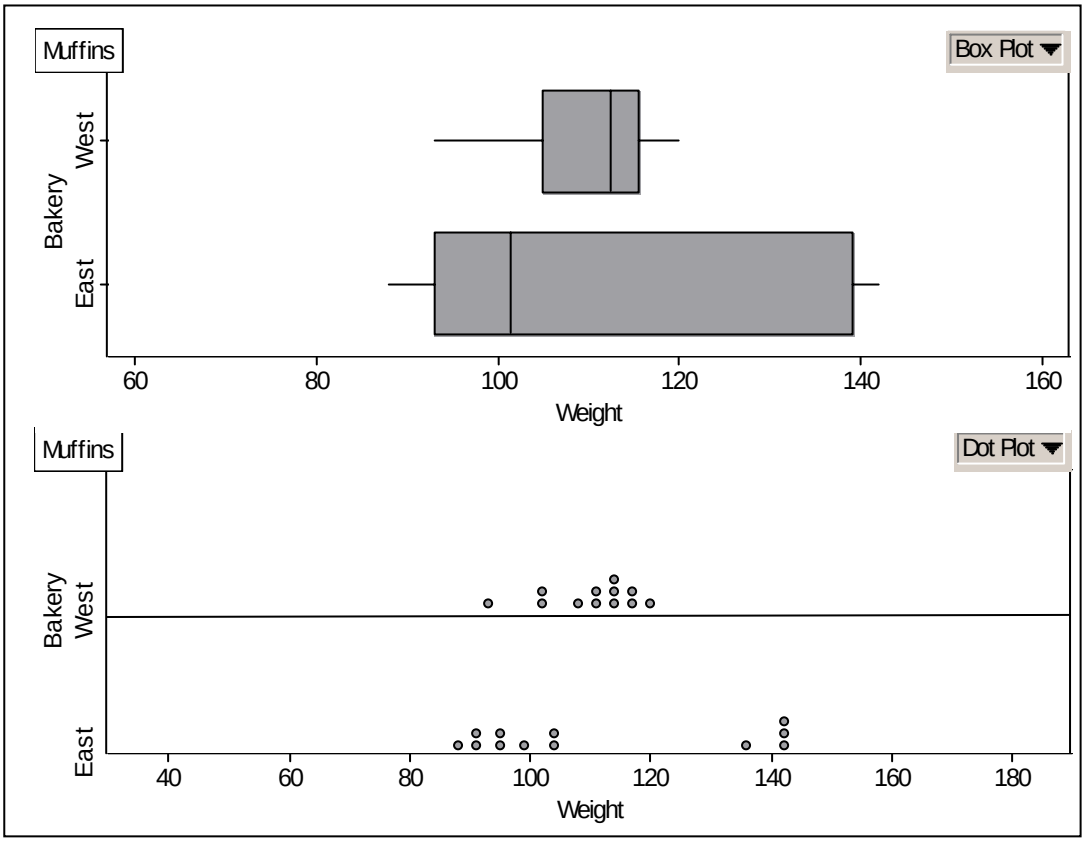


Figure 31 – PostInterview Q9 "Muffin Weights"

More importantly, the quality of his reasoning increased in sophistication in the PostInterview task:

GP: [Post Q9] Well, you see that the West End Bakery has a lot more consistency in their...the way they make their muffins.
I: How does the data show you that?
GP: Well, you can just look at the boxplot here...The middle 50% is a lot shorter than the middle 50% of the East End Bakery. You can see down here, in the dotplot, that there's quite a bit of difference where the dots are, little groupings where the dots are. The West End Bakery is closer together.

GP's comments really focus on spread, especially the way he notices the "groupings" of the dotplot and the interquartile range of the boxplot. He reasons from both types of graphs for both bakeries in making his conclusion.

Interpreting: Many of the above responses have already showed how GP defined variation in terms of having difference, or being "random." Ranges were also a part of his definition, meaning a wider range corresponded with more "randomness." His response in PreInterview Q13 showed some of his ideas about variation as he reasoned about the "Likelier Graph?" :

GP: Uhhh, Group A is...more spread out.
I: Okay, what shows you that?
GP: You have some out here, like 13, 30...more randomness to it, I guess. This [Group B] is more bunched up. Probably Group A would be more expected.
I: Any reason for that?
GP: Just more random. I dunno

Later, GP noticed that "the middle [of Group A] is more random too," and he seemed to use the term "random" to describe spread as well as range. On the PostInterview he still used the term "random" in reference to variation he was noticing, but he also broadened his definition to include other terms. For instance, on the similar "Likelier Graph?" question on the PostInterview (Q13), he said about Class A and B's graphs:

GP: Well, Class A is a lot more compact, less range. Class B has a wider range, a lot more different variations...

GP's use of the word "random" featured prominently in an early PreInterview exchange in which he focused on the physical nature of the candy mixing. I asked him why he thought results for "Several Samples" of the Small Jar would probably not be the same each time:

GP: [Pre Q1b] Because, I mean, you pour in the different candies, and [they're] all mixed up, and you might grab in a different place or pick different ones, and...it's kind of random.
I: You're saying 'random'. What do you mean by that?
GP: Random being...You can grab from so many different places in the jar. So your hand can go this way, this way, this way [He mimics with his hands how he'd hold the jar and grab in different ways] this way, this way... And then, while you reach in, the candies move, all over the place, and so... Your hand creates randomness.

Since the above exchange was during the PreInterview, it was my first chance to see how really animated GP was. As far as his reasoning, GP clearly had a notion about the physical causes of variation. Most interestingly, he seemed to be saying that if only those candies would stop shifting around, and if one could grab the same way each time, then variation would be minimized.

He made fewer references to the nature of the candy mixing in the PostInterview than he had in the PreInterview. Instead, the bigger numbers of candies in the Large Jar of the PostInterview clearly caught GP's attention. On Post Q1a, when I asked about results for "One Sample", his first reaction was:

GP: [Post Q1a] Since there's more...you're more likely to get more of the, I don't know, extreme numbers...you know, the higher end and the very few, since there's more choices.

Thus, the sheer quantity of candies is, for GP, influential on expectation and variation of results. He repeated the theme concerning sample or population size a few times in the PostInterview. When comparing results of forty samples each from the Small and Large Jar on PostInterview Q5, GP was very clear that the Large Jar should have a wider range than the Small Jar:

GP: See, the thing is, this [Small Jar] is a wider range, it seems like, than this [Large Jar] … and this [Large Jar] should have the wider range.
I: So the Class B in your mind, should have the larger range?
GP: Yeah, because you're grabbing from a larger group, and you know, [Class B] should have a larger range than that [Class A], the smaller container.

Aside from the fact that GP's interpretation is contrary to what statistical theory suggests, the point is that GP clearly saw that the sizes of the jars influenced the expected ranges.

As a last example of GP's *interpretation* of variation, I knew from his PreSurvey responses that he thought more samples might increase the range of results. On the PostInterview, he added the idea that more samples also meant getting an average of results closer to the expected value:

GP: [Post Q1a "One Sample" of the Large Jar] The more you grab, the closer you'll get to 60 and 40 being...your 'grab', the more and more you grab...
GP: [Post Q10b "Compare Samples" of the spinner] It's all in the spin, you know. The more you spin it, though, you're gonna get closer to 25

His first of the above two comments makes it seems as if he was saying a single sample result will be closer to the expected value. However, I think based on some of the class activities where we kept a cumulative average of multiple samples, and based on some of GP's other comments, he meant that the average of all samples would move closer to the expected value.

In conclusion, GP developed some appreciation for how unlikely some extreme values were in sampling and probability situations, and he expects variation in results for multiple samples. My impression is that GP is a very visual and kinetic learner, which would be consistent with the way he instinctively seemed to relate to

graphs and also was concerned with physical causes of variation. In the classroom setting, GP was the only student who exploited the slight tactile difference in the chips we used for sampling. Thus, he was able to select all green chips, for example, and knew about skewing results due to physical manipulation. He reasoned fairly well with graphs, and incorporated new terminology and graph types into his PostInterview reasoning. He already attended to average and range in the PreInterview, and in the PostInterview he had increased attention to shape and spread. One concept that would be good to explore further with someone like GP is the effect of the sample and population size on expectation and variation, since GP showed some naive understanding of that effect on the PostInterview.

The Case of EM

EM was quite willing to share what she knew and didn't know, and she needed very little prompting to voice her thoughts during both interviews. She had taken MET 1 the prior quarter with Steve. Although she was clear about not having had any prior courses in probability or statistics, her responses on the PreSurvey and PreInterview reflected her familiarity and comfort with mathematical ideas. Looking ahead to the material we would be doing in class, she said she was "open to it" and "interested to learn more". Her response at the start of the quarter to what variation meant was that it had something to do with when "there is a pattern and something changes in the pattern," and she gave as an example the time in the morning when her dog awoke each day.

Summary: EM *expected* some reasonable results in both interviews, but her sense of appropriate ranges and extreme values was not always consistent from Pre to Post. On the one hand, she predicted some reasonable ranges for "Six Samples" in both interviews. On the other hand, she had some poorer evaluations in "Comparing Lists" on the PostInterview than she had in the Pre. The biggest change for EM in this aspect was in her reliance on experience. Although she referred to her *own* instinct and experience on one PreInterview question, she made many references to *class* experience on the PostInterview. It was clear that the class interventions had made a significant impact on her reasoning.

When considering *displays* of variation, EM used more elements of the distribution in discussing graphs on the PostInterview than she had on the Pre. However, EM paid minimal attention to averages when evaluating graphs on both interviews, especially on both "MAX Wait-Times" (Pre Q8) and "Muffin Weights" (Post Q9). Instead, she talked in both interviews about how a train or bakery was consistent, and her reasoning focused mainly on the range.

There were two significant differences for EM in the *interpreting* aspect. The first difference was that while she hardly ever volunteered thoughts about the influence of the number of samples on expectation and variation in the PreInterview, she made reasonable several observations in the PostInterview. The second difference was how the PreSurvey and PreInterview picture I got of EM's perception of randomness and variation was that enough math or science could provide her with correct predictions, but she didn't convey any of those perceptions on the

PostInterview. A significant way in which EM was consistent on both interviews was in her attention to physical causes of variation.

Expecting: On PreInterview Q1a, EM guessed 6 reds for the results of "One Sample" from the Small Jar, and she reasoned proportionally. On the identical question from the PreSurvey, she had written "5 or 6", so she did have a sense that results from "One Sample" might not necessarily be the expected value. Her answers on the PostInterview "One Sample" questions were ranges:

> EM: [Post Q1a "One Sample" of the Large Jar] I think I would get, you know, somewhere in between 50 to 70 reds.
> EM: [Post Q10a "One Sample" of the spinner] I think it will land there somewhere close to 50% of the time...I think it will be probably between 40 and 60% of the time. Out of 50 spins, somewhere between – What would that be? Between 20 and 30 spins.

EM's adeptness at calculating 40% and 60% of 50 spins showed her proficiency in proportional reasoning, and she exhibited similar mathematical fluency in the PreSurvey and PreInterview. The more important point in the above two responses is EM's emphasis on range expectations in the PostInterview.

Her range for "Six Samples" from the Small Jar was adequate ("2, 5, 5, 6, 6, 8") on the PreInterview and also for the Large Jar on PostInterview ("54, 58, 60, 62, 65, 70"). Elsewhere on the Pre and PostInterview, when she had to list choices her ranges were also plausible. For instance, on Pre Q11 her results for "Six Samples" of sixty tosses of a die were "5, 8, 10, 12, 13, 15". On PostInterview Q10c, she predicted results for "Six Samples" of the spinner as "18, 20, 23, 24, 28, 32".

EM showed a peculiar inconsistency when "Comparing Lists" on the two interviews. In PreInterview Q2, she correctly thought list (i) – "7, 9, 7, 6, 8, 7" – was

too high and that list (iv) – "2, 5, 4, 3, 6, 4" – was too low. For list (v) – "3, 10, 9, 2, 1, 5" – she did not think the upper and lower extremes were likely. However, on PostInterview Q2, EM tended to favor high lists. She acknowledged that list (i) – "72, 91, 74, 63, 81,78" – was high overall and said she liked it anyway:

> EM: [Post Q2] I like the first one. Choice (i) has good variation, the numbers are also above 60, which I think is more likely to happen than below. I like that there are some 70s, I think the 91 is a little out there, but... Occasionally I think you are going to pull... [Something high]

When EM said list (i) had good variation, she meant that the numbers were all different from each other. I think EM had an unrealistic sense of just how unlikely 91 reds really is for sampling from the Large Jar. She seemed more cautious about the low list (iv) – "53, 41, 34, 60, 46, 52" – for which she thought both 34 reds and 41 reds were low. On the "Compare Lists" question for the spinner (Post Q11), EM again favored the high list (i) – "38, 43, 36, 26, 41, 33" – acknowledging that "it was definitely higher than the 50%." She cited class experience as a reason why she thought the results of list (i) were likely, and then she also went ahead and accepted the low list (iv) – "15, 19, 11, 25, 21, 23":

> EM: [Post Q11] I picked (iv) because I figured: If I'm gonna [go] high on number (i), I could see it going low, where maybe you would only get black 11 out of 50 times.

I was uncertain which class experience led EM to think that results might be generally high or low, since I don't recall any group presentations of such results. In class, we tended to aggregate results from multiple samples and those results were always on both sides of the expected value. However, the aggregate results encompassed more than six samples, so it could have happened that EM encountered runs of six samples

that reflected the high and low lists.

In any case, EM was certainly influenced by experience, and in the PreInterview she cited experience as a reason for why she knew she wouldn't get all tens for each face of the die in sixty tosses. At many other times in both interviews, EM stressed how results from multiple samples probably would not repeat each time:

> EM: [Pre Q2 "Compare Lists"] I don't think you're always going to pull 6
> EM: [Pre Q11 "Six Samples"] I'd be surprised if it came to be the exact same number as...before. I mean, yeah, It could happen, but I'd be surprised.
> EM: [Post Q1c "Six Samples"] I don't think they'll pull 60 every time
> EM: [Post Q10a "One Sample"] I don't think it will always be 50% of the time

On PreInterview Q9 ("One Sample" of the die toss) she put "8, 9, 10, 10, 11, 12" for each face, although she said "there's no reason not to get 10 of each." In fact, experience was EM's reason for not putting all tens, because later she said she was using "instinct, thinking about when I've played games, and how often sixes came up, and how often fours..." EM had naive reasoning about how often results might repeat when tossing a fair die, because earlier in Q9 she said:

> EM: [Pre Q9] Pretty often you're going to get, I don't know, every six or seven times [tosses of the fair die] I think you're going to roll the same number, again...

It is hard to imagine that EM actually based her response above on experience, and instead I suspect that proportional reasoning was influencing her thinking, based on her response above. Her informal experience also clearly influenced the list of numbers she provided in Q9, and she noted: "I just know that when I play games I usually don't get ones, so I made that one smaller."

On the PostInterview, EM made many comments about class experience. For example, on "One Sample" from the Large Jar, she reasoned that

> EM: [Post Q1a] From what we've done in class, when we pulled handfuls before, we can see that the numbers generally center around the same kind of percentage as there are red to yellow, so 60% red, 40% yellow, so somewhere between 50 and 70%

I liked how she included distributional language of how results would "center around" the expected value, and she appealed to a range of results. She based her response on what she saw in class, and she reasoned similarly in "Comparing Lists" on Post Q2:

> EM: [Post Q2, List (iii) – "60, 60, 60, 60, 60, 60"] Just from what we've done in class, I never pulled a number the same time, six times in a row
> EM: [Post Q2, List (vi) – "30, 10, 90, 20, 60, 50"] On choice (vi), I don't like, because it's too low, the 10 is too low. From what we saw in class, you know, it took I think something like 500 tries before we got so low a number.

EM's comment about the "500 tries" meant that she was recalling the computer simulation, since our hand-drawn samples usually totaled less than 300 for the entire class. For "Six Samples" of the spinner (Post Q10c), she knew "after having done it in class" that results would be "generally concentrated" around 25 blacks, which was a reasonable expectation. Experience was also her reason for why she liked the high list (i) on "Compare Lists" for the spinner (Post Q11), since she claimed that "I've seen it happen, so I liked it." Finally, using identical reasoning as she had for pulling all 60s from the Large Jar, she didn't like all 25s for the spinner, saying "I didn't pick 25, 25, 25...[List (iii)] – Although it's possible, I just haven't seen it happen, so I didn't pick that". I think it is good that the class experiences had left such an impression on EM, but I also think her thinking shows the dangers of relying too much

on experience. She often seemed to argue more on the basis of just what she had or had not seen, and less on the basis of what was more or less likely regardless of what she had seen.

Displaying: Throughout the interviews, surveys, and class interactions EM made it clear that she knew about averages and how to use them, but when talking about graphs EM tended to say more about other elements of the distribution besides the average. For example, in her evaluations of the "Graph: 30" and "Graph: 300" questions on both interviews (Q3 and Q4), EM paid more attention to range and spread than to average. She also paid minimal attention to average on the "Compare Graphs" questions (Q7), instead focusing more of her comments on the rounding procedures for each graph and spread of the data.

On PreInterview Q8, part of the interview script asked for her comparison of "MAX Wait-Times" when the averages were the same:

> I: [Pre Q8] A student in class argues that there really is no difference in the wait-times because the averages are the same. What would you say to this student...?
>
> EM: [Laughs] Ohhh. I know that the average says that, but you also have two ends of the large spectrum on the Westbound train, and a shorter spectrum on the Eastbound train. So, even though the average wait time may not differ, the amount of wait time could be a lot less, or it could be a lot more on the Westbound.

Aside from the time when I specifically directed her attention to the average, her other comments about the trains primarily concerned reliability and range:

> EM: I would say that the Westbound trains are less consistent in their wait-times. There's more variance. So, you can be waiting 7 minutes, and you can be waiting 14 minutes. And then, the Eastbound trains are pretty consistent, anywhere from 8 and a half to 11 and a half minutes. And nothing falling outside of that, so.

For EM, the term "variance" in her response above refers to the wider range on the Westbound train. On the PostInterview "Muffin Weights" Q9, the West End Bakery had the narrower range, and EM's reasoning was similar to that on the PreInterview:

> I: [Post Q9] How do you think these bakeries compare to one another, in terms of the muffins?
> EM: If I wanted to go to a bakery where I had a good sense of what I was going to get, they were more consistent in the weight of the muffin, I would go to the West bakery…
> I: Oh, why is that?
> EM: Because I can see from both the boxplot and the dotplot that they are more consistent in their weights. Their weights are concentrated between 93 and 120, whereas in the East End, you have – sometime you might get an 88 gram muffin, but you could get all the way up to 142. So if I was a big muffin eater, and I wanted to take my risk that I would get a nice weighty muffin, I would try the East bakery.

As in the PreInterview, EM again paid most of her attention to the range. She reasoned both from the boxplot and the dotplot, considered spread as she talks about how data is "concentrated", and also included "risk" as a part of her decision-making process. She knew the dotplot gave more detailed information, and she explicitly tied the consistency of the West End Bakery to its narrower range:

> EM: Again, I like the dotplot just 'cause I can see exactly where each muffin's weight fell, although just glancing at the boxplot, I can see that the West bakery is more consistent because the span is smaller… or the range. I'm sorry, the range is smaller, and… That is more consistent, then.

A good example illustrating EM's overall increased sophistication in the PostInterview when reasoning about displays of variation is found in her answers to the "Likelier Graph?" questions. In PreInterview Q13, she mostly focused on ranges, and for Group B she noted:

EM: [Pre Q13] Group B is all...The number of times is all right in the middle of the graph, 17 to 23...Yeah, that would be exactly, really close to one-third of the time they landed on black. Which is what I guess would normally happen.

At first, EM seemed comfortable with the range of 17 to 23 on Group B, and she liked the spread of data being "close" to the expected value of 20 blacks, but she concluded that "Group A is more what I would expect." Her main reason seemed to be because she liked the wider range of Group A:

EM: There was the rare times when it [Group A] dropped less than 16, and way above 24. I can also see why that would happen as well. Where occasionally...the few that are way off the charts, you know, there's a 13...and then 30 times it landed on black.

She was clearly comfortable with the range for Group A, which had the graph reflecting actual data. Since she didn't explicitly bring up the average or the shape, I asked her how she felt about the fact that Group A only hit the expected value of 20 blacks one time out of twenty samples, and she countered: "But they got around 20: 21, 18, 19, somewhere in the... around one-third of the time."

Thus, EM's analysis on Pre Q13 was reasonable, but what I noticed in the similar question on the PostInterview was that her response took into account a better synthesis of the elements of the distribution. About Class A, she said:

EM: [Post Q13] For Class A, it's all – The numbers are ONLY between 20 – it looks like 22, and 28. Yeah, 22 and 28. And it's kind of, almost like a pyramid, with just a little drop off, after the 25, so it's shaped like a pyramid...it's all concentrated around the 50%

Thus, she included average, range, shape, and spread in her analysis of Class A, which she correctly thought was likelier to be fake: "This, to me, doesn't look right, that looks like somebody made that up." Her analysis of Class B was similarly rich in

detail:

> EM: Class B's are more spread out...your mode being 27, and you know, a lot – Several of the sets were within 24 and 28, and that's what I would guess. You have a 16, and you have a 34, and there is some variation in the numbers, and that seems to be more accurate because sometimes you CAN get as low as a 16, and sometimes you can get as high as 34...25 isn't the tallest number, so...heh heh

EM appeals to a subrange of 24 to 28 within which most of the data is clustered, and she also takes note of the extreme values, which are not unreasonable to her. She also doesn't mind having the mode be somewhere other than the expected value.

Interpreting: EM was the case who made the most references to class experience in the interviews, and she was also the case who made repeated mention of how she might answer if only she knew enough mathematics. However, she only talked about having enough evidence, or having a formula, on the PreInterview. It seems that her initial perception of variation was that she could make correct conclusions or inferences only with the proper knowledge. On the PreSurvey, for "Several Samples" and "Six Samples" (Q1b and Q1c on the PreSurvey were identical to those on the PreInterview) she wrote: "Sorry, but I don't know how to calculate these answers. I'm just going off instinct. No formula, just guessing." Here is a similar response from her PreInterview:

> EM: [Pre Q1c "Six Samples" of the Small Jar] Ohhh...I don't...I don't know... [Big sigh] Well, I think, like, I don't know...I don't have any set computation, but I think it's somewhere around 6

I included the entire transcript of her response to show how she wrestled with the question. Along with the effects on her perception – how she thought that maybe a "computation" would help her figure out results in the face of variation – EM also

exhibited the effects of variation on her decisions for "Six Samples". On the die-tossing questions of the PreInterview, EM prefaced her comments about having played games by saying "I don't have a calculation, but in my head, if I threw it, I think I'd see the same number at least once every seven times." She used instinct and experience in considering results from tossing the die because she lacked "any scientific evidence for that...or mathematical evidence."

Another major change for EM was in her sense of how the number of samples influenced expectations and variation. She only made one reference to the number of samples in the PreInterview, but made more then several such references on the PostInterview. Her thinking was that more samples would widen the range of results, and sometimes she used the inverse of this concept, meaning that less samples could have a narrower range. For example, when justifying her (reasonable) list of results for "Six Samples" of the Large Jar on Post Q1c, she stressed that "you're only pulling six times." A comparison of Pre and Post responses to the "Graph:30" and "Graph:300" questions exemplifies her attention to the number of samples. In the PreInterview, EM didn't think "Graph:30" was real because:

> EM: [Pre Q3] I think, when you pull 30 times, you're going to have even more variety of times that you pull reds, and I can't believe that not once out of 30 times would they pull... no less than 5 reds.
> I: Ok, so it's the "no less than 5" that bothers you?
> EM: Yeah, it IS the "no less than 5" that bothers me...AND a little bit about the no more than 7. I think sometimes that you might pull 8 or 9, at least ONCE.

So, EM thought that 30 samples was enough to guarantee her wider range than 5 to 7 reds when sampling from the Small Jar. For samples from the Large Jar (Post Q3),

EM similarly thought about "Graph:30" that "maybe you would have a few over 70, or maybe one lower than 50, in thirty pulls." I commented that her own choices for "Six Samples" from the Large Jar (Post Q1c) had been between 50 and 70 red, and she countered that

> EM: That was only out of six pulls. And six, I like that idea, but with thirty pulls, I think you're going to have more – chance for the numbers to be a little...more spread out.

She repeated her conviction about more samples having results that were more spread out later in analysing "Graph:300" on the PostInterview. Also, she considered 45 blacks for the spinner in Post Q11 "Compare Lists" to be too high for list (vi), but it could happen "if you did 5000 sets."

A final comparison of EM's thinking for the *interpreting* aspect concerns her attention to physical causes of variation on both interviews. For example, in sampling from the Small Jar in the PreInterview, consider her following responses:

> EM: [Pre Q1b "Several Samples"] Well, I mean, some yellows might get...extra yellows might get mixed in there. It's kind of the draw I guess, how many fall into your hand
> EM: [Pre Q1c "Six Samples"] Occasionally, maybe some yellows got pushed over to the side, so you'll pull more yellow
> EM: [Pre Q2 "Compare Lists"] In case, you know, some more yellow have gone into the batch in the jar...I think that in some places, you're not always gonna have a red/yellow red/yellow...In some places, they'll be yellows that have collected together.

EM created a very vivid picture of what she anticipated. She pictured possibly grabbing more yellows in her handful of ten candies because her hand might hit a pocket of overwhelmingly yellow candies. Her attention to physical causes was not limited to sampling, and in considering "Six Samples" of the die toss for Pre Q11, she

claimed:

> EM: [Pre Q11] I just don't think you're likely to get the same answer every time. There's no way you could do that unless you know how to drop the dice, or something.

Her implication was that someone could "know how to drop the dice" and thus get repeated results. In the PostInterview, she was concerned over whether the spinner was working properly on "Six Samples" of the spinner (Q10c), and her comments about the Large Jar sampling sounded very much like what she had said for the Small Jar:

> EM: [Post Q1b "Several Samples"] A big bunch of yellows might be there and that's where you reach. You're shaking it all around, but...that's where your hand goes, and so maybe you pulled some more yellows...

She mentioned physical causes less in the PostInterview than she had in the Pre, but it was clear that she was concerned about where those yellow candies were in the jar on both interviews.

In conclusion, class experience obviously had an effect on what EM expected. Although she frequently gave good ranges for results of multiple samples, she also allowed for some fairly unreasonable results because experience supposedly suggested to her that such results could occur or had actually occurred. Also, she knew that results for multiple samples would not necessarily be the same each time, but she had an interesting expectation of a pattern of results for the PreInterview die tossing question. She did not often mention average when considering displays of variation, but otherwise improved in her ability to talk and reason distributionally about graphs. Making decisions on the basis of where she perceived more reliability was important to her on both interviews. She had increased attention to how more samples would

broaden the range of results in the PostInterview, and on both interviews she seemed concerned about physical causes of variation. What stood out for me in EM's interpretations of variation was how she repeatedly seemed almost apologetic about not having the right math to figure things out on the PreSurvey and PostSurvey, but stopped reasoning along those lines in the PostInterview. The change in her thinking seemed to reflect the impact of the class activities. That is, regardless of the theoretical expectations, for which some formula were derived or provided in class, actual results still vary.

There is a nice connection back to something that EM had first put on the PreSurvey in about what was the meaning of "random" to her. She wrote that it meant "no rhyme or reason – There is no formula." Randomness and variation together make up the Janus of stochastics: Randomness looks to the domain of probability and variation looks to the domain of statistics, but they are still two faces of the same coin. For EM at the start of the MET 2 course, she saw the variation in the outcomes of random events and wanted a formula. After multiple experiences with probability and statistics in class, she no longer mentioned wanting a "set computation," suggestive perhaps of a more accommodating or accepting attitude towards variation.

The Case of JM

Although JM was very adept at sharing his serious thoughts about variation, he also flavored his speech with levity, such as when he mentioned bringing his "triple-beam balance" to the two bakeries in the "Muffin Weights" problem, or if the person

doing trials at the spinner "wasn't drinking the night before." He seemed quite at ease during both interviews, was quick at expressing thoughts when he was certain, but pensive when he wanted to mull over a situation for which he was uncertain. JM had taken MET 1 the prior quarter with Steve, and when asked on the PreSurvey if he had taken any prior courses in probability and statistics, JM wrote "no, not really. A little sociology," which I assumed might have included a small amount of statistics. He described his own attitude going into the course in a positive way, saying it "sounds great, looking forward to it." Each of JM's interviews lasted longer than any of the other cases. A taste of how JM tended to be more expansive in his responses came early in the PreSurvey, when he gave a more protracted definition of variation as "something that fluctuates and is somewhat unpredictable. There is variety or differences." He then went on to give four separate examples of things that vary: "The weather, people's attitudes, the shapes of rocks, snowflakes."

Summary: JM was a strong proportional reasoner who shifted from having more emphasis on centers in his *expectations* in the PreInterview to having more emphasis on ranges in the Post. The biggest change for JM was that he seemed to contradict himself within the PreInterview but not within the Post. The two areas of contradiction for JM in the PreInterview concerned his sense of possibilities and likelihoods for extreme values and also for repeated values.

With *displays* of variation, JM demonstrated a very comprehensive and consistent ability to make sense of graphs on both interviews. However, even though he carefully analyzed every graph, on the questions that asked if the graphs were real

or made up JM seldom had confidence in making a decision. On such questions he tended to draw his own idea of what he thought the graphs should look really like. He was very consistent in using centers, ranges, shapes and spreads of distributions on both interviews.

There were two big shifts for JM in *interpreting* variation. First, he said much about physical causes of variation in the PreSurvey and PreInterview, and he said relatively little on the PostInterview. Secondly, on the PostInterview his ideas about the effect of more samples were more comprehensive than on the Pre.

Expecting: JM clearly knew how to calculate the expected value in figuring results for the "One Sample" questions, and in PreInterview Q1a his answer reflected his sense of proportion:

> JM: [Pre Q1a] I'd say, basically, six out of ten. There's a chance of six out ten.
> I: Why do you think that?
> JM: Well, because the ratio is 6 red for 4 yellow, for every ten there's 60%...

Similarly, on "One Sample" of the 60 tosses of the die in PreInterview Q9, he reasoned proportionally and focused on the expected value. He listed all tens for the faces of the die, and his justification was brief: "Well, it's one out of six." There were many other examples in the PreSurvey and PreInterview in which JM tended to emphasize a point estimate rather than talk about ranges.

In contrast, on the PostInterview all of his expectations were stated in terms of being close to or around the expected value. More importantly, his expectations on the PostInterview almost always included a range of possible results. He frequently

used the phrase "plus or minus" or some version of that idea to convey his range expectations. For example, compare his above response to "One Sample" of the Small Jar in the PreInterview to his response on the similar task with the Large Jar on the PostInterview:

> JM: [Post Q1a] Um, since the mix is 60% red...I'm gonna get close to that, maybe plus or minus...I'm going to say it's going to be a good... Between, you know, 45 and 80

He later narrowed his range down to "45 to 75, somewhere around there." He said he liked list (ii) – "61, 73, 56, 69, 59, 48" – because "it's around 60, but it has a decent distribution of, looks like, 20% either way from the actual number". In other words, JM liked the range around the expected value of 60 red. Also, in his comments about samples with the spinner on PostInterview Q10, for each part of the question JM emphasized ranges in explaining either *what* he expected or *why*:

> I: [Post Q10a] How many times, out of 50 [spins] do you think the arrow might land on black?
> JM: Well, approximately 50%, but it will be, you know, plus or minus, maybe 20% of that number – Somewhere in there
> I: [Post Q10b] Oh, the results on the second set, would be...
> JM: Yeah, I think [it'd be] fairly close in the sense that it's gonna be around the...uh, 25 blacks, plus or minus that 10% or so...
> I: [Post Q10c] So, 21, 23, 25, 26, 27, 29! Why those numbers?
> JM: Well, they're close to that 50 percentile that we're looking for, plus or minus – I'm thinking, 10% or so. Actually, I'm a little high, aren't I, with the 29? But still...

JM seemed fairly flexible with his ranges. At first (Q10a) he said plus or minus "maybe 20%", and then backed down to plus or minus "10% or so" in Q10b and Q10c. What was significant to me was that JM clearly had a preference for range expectations in the PostInterview that was beyond what he indicated on the Pre. His

language of "plus or minus" mirrored what he and others in class had said when discussing predictions, particularly in talking about what was in the Unknown Mixture.

One area in which JM contradicted himself during the PreInterview concerned the possibility of extreme values. On Q1a, "One Sample" at the Small Jar, JM said "you could pick out 10 red. Or you could pick out 10 yellow." Later in the PreInterview, on Q1c, he justified his prediction of "2, 3, 4, 5, 6, 7" for "Six Samples" by saying "you know, of course, you CAN pick out ten red, or you can pick out zero red." By the time of "Graph: 300" in PreInterview Q4, JM backed away from his emphasis on the possibility of extreme values:

> JM: [Pre Q4] I can't believe that there were, um, that you could...I don't think you COULD pick up all 10 red, or all...or zero red. I think there would have to be some [yellow?]...I just think it's impossible to pick out [all reds], if they're mixed.

I did not point out JM's inconsistency to him, but I found it interesting how the following pattern seemed to emerge with JM: When JM was asked to predict results, he was careful to mention how extreme outcomes were possible, yet when he was shown purported results (particularly in graphical form) he was skeptical of the extremes. On the PostInterview, JM still emphasized the possibility of extremes, but did not contradict himself. He said, for drawing "One Sample" from the Large Jar in Q1a, "I mean, it's possible to get zero red, and it's possible to get 100 red." His most common way of talking about extreme values in the PostInterview included how those values were possible but unlikely. For example, in considering list (vi) – "30, 10, 90, 20, 60, 50" – JM commented that "when we go to extremes like that, they're highly

unlikely and to have those...It's possible." I thought that JM had moved to a somewhat more balanced view about extremes in the PostInterview. Instead of swinging between polarized opinions of what could or could not happen, he had a better sense that some outcomes, while still possible, were "highly unlikely."

The other area in which JM contradicted himself during the PreInterview concerned the likelihood of repeated results. When I asked him in Q1b if he thought he'd get the same results every time for "Several Samples", JM was quick to stress "no, no...of course not." When I asked him why he thought results would not repeat, he said "well, because it's...it's impossible." Thus, it was clear that JM did not expect results to repeat for multiple samples, and on his own choices for "Six Samples" he did not repeat any values. However, by the time of PreInterview Q2, JM liked list (iii) – "6, 6, 6, 6, 6, 6" – saying initially that "Choice (iii)'s real good, because it could be 60% every time [laughs]." JM went on to wrestle aloud with the twin notions that getting all sixes was possible but unlikely. My main surprise for JM came in PreInterview Q9, when he listed all tens for the faces of the die in "One Sample" of sixty tosses, saying his choices were "not unreasonable." When JM considered Lee's supposed results of all tens in "Who Cheated" on Q10, JM said "I don't think, even if I rolled them 60 times, I would not get 10 numbers each." As I had done with DS, I pointed out how JM had listed all tens on Q9, to which he replied:

> JM: I think it's possible, likely, because it's one out of six times, but I don't think I could roll that and that'd actually happen. I said they're REASONABLE...

JM then went back and started to change his list of all tens on Q9, but then he resolved

to leave his list of all tens in place, and continued to defend his choice. His eventual argument was that

> JM: I think there's a greater chance of it coming up the ten, than maybe another number. And to pick ANOTHER number is, it'd be just as good as maybe picking ten. Since it's [the probability] 1 out of 6, I mean, I would pick 10.

JM was willing to stick with his own list of all tens, saying his list was "possible, but not probable", and clearly wrestled with himself over how likely he really thought it would be to see repeated results in multiple samples. In other tasks on the PreInterview, and all throughout the PostInterview, JM took a more moderate view of repeated values. That is, he included in his responses both the possibility and the unlikeliness of getting the same result each time for multiple samples.

Displaying: JM used all elements of the distribution at some point in both interviews as he deliberated the questions involving graphs. While he did not demonstrate any dramatic changes in reasoning about *displays* of variation from Pre to Post, he did add some sophistication to his discussion in the PostInterview as he invoked new ideas (such as the interquartile range) that he had gained in class.

To illustrate how JM reasoned about graphs, consider his response to PreInterview Q8 about the "MAX Wait-Times." At first, he pointed out how the two trains had the same means and medians. Then he said that there was a "big difference" between the trains because:

> JM: Even though we have 2 thirteens and a fourteen or a fifteen minute on the Westbound train, that's just 3 trains out of , what? Out of ten? Ten trains? So 30% of the trains take longer than 12 minutes on the Westbound train, and then , um, you know, it looks like 100% of the trains are under 12 minutes on the Eastbound train.

I thought JM's above response showed a reasonable attempt to compare the distribution of wait-times between the trains, and he later appealed to the shorter range of the Eastbound train in declaring it to be more reliable than the Westbound train:

> JM: Well, the Eastbound's are much more reliable. They go from 8 and a half to 11 and a half minutes. So you don't have this broad…You're going from 7 minutes to 14 and a half minutes on the Westbound train, so the chances of you waiting - shorter than the Eastbound train – are … it's only 30% of the time, most of the time you're going to wait equally or more, on the Westbound.

Thus, JM noted how the Eastbound train a narrower range than the Westbound train. He also calculated what percentage of the Westbound trains had longer and shorter wait-times than the Eastbound, which to me seemed to be JM's way of getting a sense of the relative spreads of the two data sets.

In the isomorphic "Muffin Weights" task on PostInterview Q9, JM reasoned similarly to the way he had in the PreInterview. Again, he first focused on the median, saying that "obviously the west End bakery produces, on average, a bigger muffin." Then, he shifted his attention to the range and spread of the data, using the boxplots as a basis of comparison. He rightly noted that "the interquartile range – 50% of all the muffins [in the East End bakery] – exceeds the whole range of the West End bakery." Because of their wider range, JM said "I'd probably be less confident of going to the East End bakery." He then made some astute observations about the three measures of center, which were all different between the two bakeries. For example, the West had a higher median than the East, but the East had a higher mean and a higher mode than the West. JM took note of how the data was distributed along the dotplots for the two

bakeries, and concluded:

> JM: We really wouldn't want to look for the typical muffin in the mean, we'd want to look for it probably more in the median.
> I: Oh, why is that?
> JM: Well, because there's just too many...the range is...too many low-weight and high-weight muffins. That really throws off our idea of the typical muffin.

I was pleased to hear that JM had some notion of the effect of variability on the mean in context of the "Muffin Weights" task, because a comparison of the merits of different measures of center had been a part of the class curriculum.

I'll use Q13 from both interviews – the "Likelier Graph?" task – to illustrate two tendencies that JM showed when comparing and evaluating graphs to see if one graph or the other was likelier to reflect actual data. One tendency was to not have much confidence in deciding whether graphs were real or fake, and the other tendency was to sketch on the interview script to show his own idea of what results should look like. He also showed these two tendencies in "Graph: 30" and "Graph: 300" (Q3 and Q4) on both interviews. In the PreInterview Q13, JM was quick to compare ranges in talking about which was the "Likelier Graph" :

> JM: [Pre Q13] Group A certainly has wider variables [Holds hands far apart], it's gone between 13 and 30. And Group B of course, is a much tighter distribution of black [Brings hands close together].

He then noted how Group A lacked any entries at the expected value of 20 blacks, while 20 blacks was the mode for Group B. For a time, he leaned towards thinking Group B's graph was real, but he pondered both graphs for awhile before saying:

> JM: These graphs have got me stumped in the sense that…I would think that there would be more… More 20s here [Group A]… I would like to meld both of the graphs.

JM decided that, instead of having confidence in either Group A or Group B being realistic graphs, he would draw on top of both graphs to make them look as he thought they should. He changed Group A by moving the mode to the expected value, and he narrowed the range. He changed Group B by reducing the height at the mode, and widening the range. In the end he had two graphs that looked roughly the same, and they were both shaped like smooth bell curves centered at the expected value of 20 blacks.

In the PostInterview Q13 "Likelier Graph?" task, JM reasoned much as he had in the PreInterview, but his language was slightly more descriptive. I asked him to compare the two graphs, and he said:

> JM: [Post Q13] Class A has that nice, nice shape that I was looking for [He draws an inverted-"V" shape on Class A], though it might not be evenly distributed. And Class B is just all over the place, with a mode of it looks like 27...And the fact that there were no, nothing below 22 here, on Class A, or above 28...it was just too tight, in 30 sets.

JM attended to average, range, shape, and spread in his response above, and eventually he concluded that both Class A and Class B "have a pretty good chance of being made-up." He then drew on both graphs. He expanded the "tight" range of Class A, and he shifted the mode on Class B to 25 while also narrowing Class B's range. Because he had mentioned Class B being "all over the place," I was not surprised to hear that he wanted data to be "more evenly distributed." It was clear that JM could use all elements of the distribution in making his evaluations and comparisons of graphs, but he still under-appreciated just how scattered data from on 30 samples could look in the PostInterview. That is, he saw the gaps along the axis for Class B

(the possible outcomes for which no results were attained), and he was skeptical. He also was still expecting to see the mode at the expected value, even for only 30 samples.

Interpreting: JM volunteered many more causes of variation in the PreInterview than in the Post. For instance, in PreInterview Q1c ("One Sample" of the Small Jar) he suggested that one might get different results "depending on how they're mixed up." He seemed especially interested in how the individual candies might lie in the jar next to one another, and in PreInterview Q3 ("Graph: 30") he expressed his thinking as follows:

> JM: You know, I guess I'd have to see how they fit into your hand. Maybe that has a bearing on it possibly, right? And when you reach into a container, and pull them out, and if they're completely mixed, whereas one red is lying is against, or there'd be, what is it, 60%? So you'd have almost...2 reds around 1 white [He means yellow]. Maybe? Something like that?

So too did JM stress causes in the spinner scenario of PreInterview Q12. Although Pre Q12 did not have an isomorphic counterpart in the PostInterview, JM's initial response focused on the mechanics of the spinner, not the content of the task. I'll quote the entire exchange because it really shows JM's emphasis on physical causes of variation:

> JM: Um, well…I want to look at the engineering of the spinner, where do you start the spin, you know, I mean…. Do you start it in white, you know, the velocity, or the force… None of that really matters, I guess…I mean, it CAN matter of course, yeah. Well, of course, it WOULD matter, you know, I mean, you play like a game that has a spinner, and, if you're a kid, you know if you hit it just the right way, and you start it at just the right the spot, you could… there's a chance of it being in one spot are greater than in another spot.
>
> I: So this is very well-oiled spinner…Very, very fair spinner

JM: Ok, so this is a GOOD spinner. Yeah. Ok. A fair spinner. Um, yeah. And the spinner is, is flat? A flat plane? It's a fairly spun game?

Rather than being contentious, my sense was that JM's expectation of variation depended greatly on the physical apparatus and the actual performance of each trial, whether it was drawing candies from jars or using spinners. However, in the PostInterview he offered very few ideas about causes in these contexts, and it seemed that his side comment above about "none of that really matters" probably gained dominance over his thinking as we engaged in the class activities designed to show random behavior. He did mention the way the spinner was used in PostInterview Q10b ("Compare Samples"), and "getting into the rhythm of it [spinning]". However, JM volunteered much less about physical causes in the PostInterview than in the PreInterview.

Another change for JM was that he expressed better ideas in the PostInterview than in the Pre about what the influence of doing more samples would be on expectation and variation. On PreInterview Q1b, he indirectly appealed to the Law of Large Numbers when he commented about "Several Samples" that "I think on average, if you did it enough times, you probably average 6 reds". JM also suggested that more samples gave more chances to obtain the expected value in the "Compare Lists" question on PreInterview Q2, saying "in six tries, I would think that six reds would have to come up at least once or twice." More samples meant a broader range as well. In PreInterview Q3, JM felt that the 30 samples in "Graph: 30" would surely range beyond the 5 to 7 reds depicted in the graph. A significant idea JM expressed in

the PreInterview which was not expressed in the Post was how the ratio doesn't change regardless of the number of samples. On PreInterview Q9, JM argued in favor of his list of all tens for "One Sample" of sixty tosses of the die, reasoning that

> JM: You have one out of [six]...There's six sides. It's got to land on one of the six. And each one is, I guess, equal. So, after the first time, it's still one out of six. And the second time it's still one out of six. So...

He later re-iterated his emphasis on the unchanging ratio within the PreInterview.

On the PostInterview, JM didn't mention how the ratio is independent of the number of samples, but he repeated his earlier ideas and emphasized them more often. That is, more samples meant a widening range, more chances to actually attain the expected value, and a convergence of the cumulative average of results toward the expected value. Here are some examples of his responses:

> JM: [Post Q4a "Graph: 300"] In 300 pulls, I mean, it's gonna happen, you're going to pull out less than 48 reds, at least once. At LEAST once. Maybe twice, or three times, or four or something...
> JM: [Post Q10a "One Sample"] If he does it enough times, he's going to be right at that number [6 reds]
> JM: [Post Q10a "One Sample"] With 50 spins...that's fairly good sampling, or...number of trials...that would approximate the theoretical probability

JM's reference to approximating the theoretical probability mirrored what we had discussed in class, and he was even more articulate when talking about the "Likelier Graph?" of PostInterview Q13:

> JM: [Post Q13] So, the theoretical should come close to the experimental… over the long run, if we do enough trials, and have a big enough sampling of what we're doing. So once we figure out the theoretical, we go out and try to prove it experimentally, and see how close they come. And, chances are, they'll come pretty close if we do a fair number of sets.

JM also implied that graphs would have a better "look" with more samples, and about "Graph: 300", he said that:

> JM: [Post Q4] I think that the more sampling that you do, the more close to that nice 60% distribution you're gonna get. If we did 3000 pulls, it would even look , you know, better…

I did not probe JM's thinking in the latter response, but I believe he was referring to having a smoother bell-shaped curve with increasing samples.

In conclusion, although JM gave more range expectations in the PostInterview than in the Pre, I could see how much he was influenced by centers in both interviews. For example, one reason he was unwilling to identify authentic graphs as such is because he thought the expected value was what should occur the most often, in both interviews. His expectations about ranges also seemed inconsistent at times. In PreInterview Q4 ("Graph: 300"), he thought the range was too wide for 300 pulls, and in the isomorphic task on PostInterview Q4, he thought the range was too narrow. In both situations, the graphs were authentic. JM was adept at using different elements of the distribution in discussing graphs, but he in both interviews he lacked confidence in deciding if graphs were real or made-up. Class experience seemed to help him identify unlikely graphs, but not to help him argue that a graph was in fact likely. I think that physical causes of variation remained an important issue for JM in both interviews, but it seemed more on his mind in the PreInterview than in the Post. He had many reasonable notions in both interviews about the influence of more samples, and clearly reflected ideas about the Law of Large Numbers which had been brought out in class.

The Case of SP

SP was a very reflective individual, someone who really thought not only about her answers, but also how she was thinking and feeling about the questions. Her language suggested she was comfortable with a sort of metacognition, and she repeatedly talked about her instincts and feelings, often contrasting those thoughts with a logical perspective. For example, she talked about "my first instinct", and then how "there's not any super-logical reason" but "I guess that's just where my brain goes first." She clearly showed a willingness to try and explain what was going on in her mind. She volunteered information readily, telling me what would or wouldn't surprise her, for instance. Although she was an easy person to talk with, both of her interviews lasted a bit shorter than average. SP had taken Math 211 the prior quarter with Steve, and wrote on her PreSurvey that she had taken some probability or statistics course at another university four years ago. She recalled that it had been a "fun, interesting class," yet currently she said she felt "comfortable but shaky – don't remember much but I'm sure it will come back to me." She wrote that that variation meant to her "the differences between things in a group," and gave several examples: "Weight, height, hair color of a group of people."

Summary: SP's PreInterview ideas about how "Anything Could Happen" and "You Can Never Know" reflect the Outcome Approach detailed earlier in Chapter 2. The essence of the Outcome Approach can be characterized by an attempt to look only at the next outcome of a probabilistic event, and transfers to the sampling context by focusing on the results of the next sample drawn. I think that when SP said that she

"can't guess", what she really meant is that she could not guess with the a priori assurance of being correct. In other words, she could never know ahead of time what the outcome would be. Since she believed in the PreInterview that uncertainty meant "Anything Could Happen", and because she could never know ahead of time what would happen, that why making a prediction was the same to her as "saying anything." Her sense of how "Anything Could Happen" explains why her ranges were so wide in many of the PreInterview questions, and yet it also explains why she gave all tens in Pre Q9 even though she didn't really think that outcome would happen.

SP *expected* results for multiple samples to be usually be different and not repeat, and she made explicit references to the underlying theoretical ratio in the PostInterview but not in the Pre. Instead of using "median" numbers and wide ranges to express what she expected in the PreInterview, she offered reasonable ranges that were appropriately centered around the expected value in the PostInterview. The distributional elements of range, shape, and spread were evident in her responses considering *displays* of variation in both interviews, and she included more of a focus on average in the PostInterview. In her *interpretations* of variation, during the PostInterview (but not in the Pre) SP volunteered some very reasonable ways that the number of samples might influence expectation. She also showed a major shift in her thinking, moving from the idea that "Anything Could Happen" in the PreInterview to the notion that some outcomes were likelier than others in the PostInterview. Her related PreInterview theme of not knowing gave way in the PostInterview to a theme suggesting that while you may not know for sure about a given outcome, you can still make reasonable statements of expectation.

Expecting: SP established at many times throughout both interviews that she *expected* results for multiple samples to usually be different from one another. For example, in Pre Q1c ("Six Samples" of the Small Jar) she mentioned how she would "be more surprised if the same number kept showing up, as opposed to if it was just completely random." She acknowledged that repeated results were possible in "Comparing Lists" on Pre Q2, but maintained that the "6, 6, 6, 6, 6, 6" of list (iii) would cause her to "be VERY surprised." Even list (i) seemed "more unlikely" to SP, since list (i) contained three results of seven reds. Her responses in "Comparing Lists" on the PostInterview were similar to those on the Pre concerning repeated values. She liked list (ii) on Post Q11 because "there's not a lot of repetition". List (v) on Post Q11 – "24, 24, 25, 25, 26, 25" – was not favored by SP because of "too much repetition, you expect more variation."

The expectations that SP volunteered improved dramatically from Pre to PostInterviews. One area of improvement was how she gave appropriately wider ranges on the "One Sample" PostInterview questions. For example, in Pre Q1a for "One Sample" of the Small Jar, she said: "I guess instinctually I would say that it'd be somewhere in a median, like uh... 4, 5...just instinctually." In contrast, for Post Q1a she said that "One Sample" of the Large Jar should give her "somewhere between 50 and 70" red. Whereas she gave all tens for the "One Sample" of sixty tosses of the die on Pre Q9, for "One Sample" of the spinner on Post Q10a she expected "between 20 and 30" blacks. In the above examples, her expectations in the PreInterview are less reasonable, than those given in the Post. A second area of improvement for SP was

how the ranges she had for the "Six Samples" questions were appropriately narrower in the PostInterview than in the Pre. For "Six Samples" from the Small Jar (Pre Q1c), she listed "1, 2, 3, 4, 6, 8" , which is too wide. In the isomorphic question for the Large Jar (Post Q1c), she gave "49, 51, 55, 62, 65, 68", which is quite reasonable. Similarly, her choices for "Six Samples" of the die toss (Pre Q11) were "2, 5, 7, 1, 14, 20", again an unlikely list. SP had a much better list for "Six Samples" of the spinner in Post Q10c, "20, 23, 24, 26, 28, 29".

Another change for SP was that she included references to the theoretical ratio in the PostInterview but not in the Pre. For instance, she talked about expecting "median" numbers of 4 or 5 in "One Sample" of the Small Jar (Pre Q1a), and she repeated her preference for "median" numbers in response to a couple of other questions in the PreInterview:

> SP: [Pre Q1c] I just actually did like a median number, like 5...
> SP: [Pre Q2] And again, I don't know why I feel also comfortable with the median numbers, the 4, the 5, 6... for some reason. Yeah, like 4s, 5s, and 6, to me, is somewhere in the middle

It wasn't clear to me at that point in the PreInterview if SP even knew that the expected value for the Small Jar sampling was 6 reds. Even in the "One Sample" of sixty tosses of the die, she listed all tens but never explicitly said anything about the probability for any face being one out of six. Instead, she talked about giving all the faces "an even chance" and how she wanted to make sure her choices added up to 60. SP never articulated a single fraction or ratio anywhere in the PreInterview, showing

what I think was an under-attention to the expected value. In the PostInterview, she made it clear that she had considered the ratio in making her choices. For example, here is what she said in "Comparing Lists" on Post Q2:

> SP: There's also that 400 yellow in there, 600 red and 400 yellow, [and] the likelihood of just getting the 60 out of 100, which is like the perfect ratio or whatever, is very unlikely

An increased attention to the ratio helped SP to center her ranges in the PostInterview, such as when she had put "between 20 and 30" for "One Sample" of the spinner in Q10a, saying:

> SP: Umm, because of that 50 to 50 ratio, or chance of getting black, and chance of getting white – And so, out of 50 times, half of 50 is 25, and so that would be the, sort of – expected ratio. Not expected, but the – Theoretical ratio [Laughs] And uh, the between 20 and 30 would take into account the, the actual practice of spinning it…

The increased attention to expected values and improved sense of range that SP had in the PostInterview also allowed her to make better choices in the "Compare Lists" questions. In the PreInterview, she never commented on list (i) being generally too high, although she did feel the result of 9 reds was too extreme. Similarly, she liked list (iv) because of the "median" numbers, and never pointed out how the entire list was generally too low. She picked list (v) – "3, 10, 9, 2, 1, 5" – as her favorite because she liked the "huge...variety of numbers." She added that "Even though I said I would be surprised by 10... I feel like that [list (v)] covers a wider range . It has some high, and some really low." However, in the PostInterview, she correctly pointed out how list (i) was high overall and how list (iv) was low overall. She reasoned that

> SP: [Post Q2] 'Cause again – The perfect ratio's you would get 60 red, 40 yellow, and so...I guess you would want to go maybe 10 above or below that? Maybe more...for a variety of answers

At the end of her analysis, she picked list (ii) as her favorite on Post Q2, which was the most reasonable choice (and consistent with her own reasoning). List (ii) was also SP's favorite choice (and most reasonable) in Post Q11, the "Compare Lists" problem for the spinner. When I asked her why she liked list (ii), she said "they fall in that nice little range of mine, between 20 and 30, but with a few just going a little below and a little above, which I like." She also liked how list (ii) had no repeated values, and noted how the list had "a lot of variation." Thus, her final choices in "Comparing Lists" were better on the PostInterview than on the Pre.

Displaying: SP used range, shape, and spread in comparing and evaluating graphs during both interviews. The biggest change from Pre to Post was that she showed an increased attention to averages in the PostInterview. Also, while she correctly discerned real from fake graphs in both interviews, she seemed more confident of herself in the PostInterview.

I'll first illustrate her reasoning about *displays* of variation using her responses to the "Graph: 30" and "Graph: 300" questions on both interviews. In Pre Q3, at first SP spent some time wondering if she could have any confidence in knowing if the graph was real or fake. Eventually she thought that "Graph: 30" was made up because she did not "feel comfortable" with the shape of the graph, and we then had the following exchange:

SP: [Pre Q3] I like the wider range of things, I feel like that's more likely to happen. If, you know, to have it more random and this [graph] seems really less random.
I: Oh. What makes it seem less random to you?
SP: Because they're all 5, 6, and 7s. And three numbers in a row and...all clustered

SP thought "Graph: 300" was "more realistic", and she liked it because "it's spread out, there's a little bit of everything, and then...most of them are somewhere in the middle..." She had a slight reservation about "how perfect" the graph was, and when I asked her what she meant she said:

> SP: [Pre Q4] As in, it, you know, it's like this perfect curve. Whereas I don't know if something's randomly being chosen, that you can get this perfect curve. [Traces the shape with her finger]

She continued to emphasize shape as I asked her to compare "Graph: 30" to "Graph: 300", saying that the former had "a more extreme curve" while the latter had "a more gentle, gradual curve." I thought she reasoned well in the PreInterview, but what she added to her reasoning in the PostInterview was an explicit attention to the center of the distribution. Her entire response to Post Q3 is an excellent example of the overall improved caliber of communication:

> SP: [Post Q3] They seem like they could be the actual results.
> I: What convinces you, or what is your reasoning?
> SP: Because they fall into that sort of theory, that you're going to have the most around 60, because of that perfect, that 60 to 40 sort of ratio, or 40 to 60, whichever. And so, you sort of that happening, and then they fall out in about a range of plus 10 , minus 10. So it makes sense. But they – It's random enough so that it's not like this perfect bell-curve, so it seems like more of a realistic situation because it's not perfect.

She also mentioned how "Graph:30" was more "scattered" and that was why it was "not perfect" to her. She did a good job of synthesizing several elements of the distribution in her response, including an explicit mention of the average of 60 reds. SP also pointed out the mode of "Graph: 300" in Post Q4, and she commented on the

graph's shape, spread, and range in deciding the graph reflected genuine data:

> SP: [Post Q4] But every once in a while it would happen that you would get, somewhere, like 48 or something like that, or a 73… And , yeah, it's starting to conform more to that bell-curvish, where you're getting mostly results from, like, 58 to 62 [The modal category] which you would think to happen, and again, it spreads out from there, along the range

I particularly liked the language she used when she pointed out similarities between "Graph: 30" and "Graph: 300" in the PostInterview, because it showed her attention to the way the data was distributed:

> SP: Yeah, well, the bulk [of the data for "Graph: 30"] is still sort of in this little area here [She circles her finger around 60 red], and of course it's a little more scattered...And it does the same thing where it goes out almost at the same distance from that center 60.

SP conveyed a good sense of variation away from the mean in her latter response.

A second illustration of SP's reasoning about *displays* of variation comes from the "Likelier Graph?" questions on both interviews. Again, as in the "Graph: 30" and "Graph: 300" questions, SP demonstrated that she had some good ideas about graphs in the PreInterview which she improved upon in the Post. In PreInterview Q13, SP correctly identified Group A as likelier to be the authentic graph. She claimed that Group A had "greater variation" than Group B because Group A had a wider range:

> SP: [Pre Q13] It's more spread out. It [Group A] goes from, the lowest is 13, and it goes up to 30. This one [Group B] is clustered within 17 to 23

In the isomorphic PostInterview Q13, she again correctly identified the authentic graph (Class B), and at first she used an argument based on shape which sounded very much like what she had said at the end of PreInterview Q4:

SP: [Post Q13] Class A is more like a drastic bell [Traces a bell curve], and this [Class B] is a more gentle bell [Traces broader bell curve]

SP went on to say how Class B had "more variation, is more spread out", while Class A was "more compact...and the range is really short." She explicitly mentioned the average (something she hadn't done in the PreInterview) when she pointed out that Class A had "not a lot of variation, and all sort of centered around that theoretical 25." SP also continued to refer to where the "bulk" of the data was for both Class A and for Class B, a term which she used to indicate where she saw data clustered.

Overall, the inclusion of the references to the average in the PostInterview was the biggest difference in SP's responses about *displays* of variation. She consistently reasoned well in the other questions involving graphs. For example, in the "Compare Graphs" questions for both interviews, SP felt that coarser rounding produced graphs that masked variation more than the graphs using finer rounding. In "MAX Wait-Times" and "Muffin Weights", she used both range and spread to help her identify the more consistent train or bakery. Throughout both interviews, she used many different descriptive terms suggestive of how data was distributed, such as "compact", "scattered", and "clustered."

Interpreting: One way that SP was fairly consistent in this aspect was how she referred to the number of candies in both interviews. In PreInterview Q1a, she said that "One Sample" of the Small Jar had "a slightly greater chance" of having more red "because there's more red than yellow." Later in the PreInterview, she repeated the same theme. In PostInterview, after she gave her range of 50 to 70 reds for "One Sample" of the Large Jar, she cautioned that she wouldn't expect "too many less,

because there're so many red in there." Later in the PostInterview she returned her focus to the number of candies used in sampling.

A significant change in how SP *interpreted* variation was that she mentioned the influence of doing more samples during the PostInterview but not during the Pre. Because I had asked questions on the PreSurvey and PostSurveys that directly invited thinking about the influence of more samples, it was clear that SP expected a wider range from an increased number of samples. However, she never volunteered any information in the PreInterview about the influence of more samples. On the PostInterview, however, she had several ideas. For example, in Post Q1a, she said "you expect a sort of, an average of 60, if you did many of these [samples]." I thought her response suggested the Law of Large Numbers, and reflected activities and discussions we had as a class. SP also thought that more samples gave more chances to attain extreme values, and she repeated this contention several times during the PostInterview. What follows are some of the different questions in which she addresses connects more samples to a widening range or more extremes:

SP: [Post Q1b "Several Samples"] But also [there'd be] some more extreme numbers, eventually

SP: [Post Q4 "Graph: 300"] They're going out a little bit even further, which which you would expect...With more pulls you would do, the more sort of outliers you would get, or the "unexpecteds" you would get...

SP: [Post Q11 "Compare Lists"] You're only spinning 50 times...But if you were spinning 100 times, maybe those numbers [extremes] would go further

The final idea SP had along this theme was that more samples influenced the shape of the graph. In "Graph: 30" she emphasized that "you only did 30 pulls, so it's going to

look a little bit more scattered." Since "Graph: 300" involved more samples, the graph "would become more...conformed to this perfect bell-curve, and that it would pull out just a little bit more." The ideas that more samples would move the cumulative average closer to the expected value, widen the overall range, and better reflect the shape of the underlying distribution were all ideas we had addressed as a class.

Another significant change was that SP repeatedly discussed in the PreInterview but not in the Post how she had difficulty guessing because "Anything Could Happen." The effects of variation on SP's perception and decisions were so pronounced in SP's PreInterview (and PreSurvey) responses that they seemed to dominate her thinking at times, and I'll highlight several examples. In PreInterview Q1a, her very first words for "One Sample" of the Small Jar were:

> SP: [Pre Q1a] My first instinct is just to say that, it could be any amount. You could have all, you could have none...[Then, later on:] But then, like, if I try to THINK about it, if I tried to think it out, then I'm thinking : It could be ANY amount, and I can't guess, you know.

She emphasized her view again in Q1b ("Several Samples"), saying that "it can be anything. Logically, that's what my brain is telling me, is it can be absolutely anything." I had already known about how SP thought results "Could be Anything" from her PreSurvey responses, but I then saw in the PreInterview how the effect on her decisions was that she did not want to make any guess at all. She continued in Pre Q1b to say:

> SP: [Pre Q1b] I think I'm just pulling out a number because I'm feeling like I should make a guess. But I really don't want to make a guess...Yeah, because I feel like it really can be anything. And so making a guess is just like…. Just saying anything.

Coming at the beginning of the PreInterview, a snapshot of SP's thinking developed which portrayed her as having difficulty predicting results because "Anything Could Happen". By the time of Pre Q3, although she thought "Graph: 30" was fake, she was "also attracted to the 'We Have No Confidence' because we REALLY can never know, because it COULD happen, there's always the chance that it COULD happen." The effects of variation on her perception decisions were particularly relevant to her reasoning on the die-tossing questions of Pre Q9, Q10, and Q11. She reiterated the following view:

> I: [Pre Q9 "One Sample"] Why do you think those numbers [All tens] are reasonable?
> SP: Because if you're forced to guess...as I've been saying, that they could be anything. So I just gave them each an even chance...I can't guess. I have trouble making guesses

SP then went on to describe how she could get the same face of the die for each of her sixty tosses, but that she'd be surprised at that outcome because "it could be any of the numbers." She declared that "I'm just giving them each an even chance, because I guess...I can never know." On Pre Q10, in discussing "Who Cheated?", SP explained how her strategy in choosing all tens on Q9 was not motivated by what might really happen, but that "when you're making a guess, I just do it that way, because you can never really guess." She went on to stress how results "Could be Anything" later in her response to Q10, Q11 ("Six Samples" of the die toss), and Q13 ("Likelier Graph?"). Other subjects had expressed similar themes about not knowing what might happen, or what could happen, or how anything could happen, but no other subject was as outspoken on these themes as SP. Thus, I was surprised that in the PostInterview, SP never expressed views about how "Anything Could Happen" or

"You Can Never Know".

The Case of RL

Of all the students in Steve's section, RL stood out in class discussions, on the research surveys, and in the interviews as having the most mathematically-oriented responses. He had a strong background in mathematics and also in philosophy, and during the interviews he would occasionally veer off on some tangent that seemed related in his mind, such as how the digits in the decimal representation of pi were randomly distributed. RL readily volunteered all kinds of information about what he thought and why, and his unprompted responses were lengthier in general than those of the other cases.

RL had taken MET 2 the prior quarter with Steve, and had taken a past college course in statistics at a different university. He also thought that both probability and statistics had been covered briefly in his own high school. Considering his own attitude at the start of MET 2, RL said: "As a future teacher, I look forward to mastering at least the basics." Again reflecting his penchant for mathematical terminology, RL's definition of what variation meant to him on the PreSurvey was "a measure of how a piece of data compares with the average of similar data." His definition corresponded well to the idea of variation from the mean, and his example of something that varies was "sea level."

Summary: Although RL clearly exhibited distributional reasoning prior to the class interventions, he had some contradictory *expectations* within both the PreSurvey and the PreInterview. For example, on some questions RL wrote or talked about expecting to see variation in results, but on some other questions he said the expected value should occur repeatedly because it was the most likely outcome for a single sample. Due to the cognitive conflict induced by the sequencing of the die-tossing questions in the PreInterview, RL began a shift in his expectations that led to a more consistent appreciation for ranges by the end of the PostInterview. RL's reduced emphasis on centers from the Pre to the Post was also accompanied by a reduction in his frequent references to mathematical computation, and an increase in his focus on distributional reasoning.

RL seemed to misidentify real versus fake graphs for different reasons in considering *displays* of variation during both interviews, but overall I think he was relying on the Representative heuristic mentioned in Chapter 2. He liked "Graph: 30" in PreInterview Q3 because of the symmetry and the center, erroneously thinking that results for 30 samples would be a fair representative of the underlying distribution. He used the same kind of reasoning in considering the "Likelier Graph?" of PreInterview Q13. The graph for Group A he incorrectly labeled as fake because it was too "wild" and not as representative of the underlying distribution as Group B. In particular, RL thought that Group A had too many extremes. In both PreInterview Q3 and Q13, the small sample sizes used are not likely to yield graphs that are very representative of the population distribution, but RL did not seem to appreciate that fact. On the other hand, in PostInterview Q4, "Graph: 300" would be expected to give a reasonable

representation of the overall distribution. RL thought that the graph did not go wide enough, however, and was suspicious of the graph's authenticity.

In his *interpretations* of variation, RL reminded himself in the PreSurvey and PreInterview how reality was different from theoretical expectations, but he didn't often give voice to those thoughts in the PostInterview. He was the most outspoken subject in terms of the influence of the number of samples, especially regarding the influence on the shape of the distribution. He knew that overall ranges would expand with an increased number of samples, yet also expressed in the PostInterview how relative ranges would tighten. He seemed to focus more in the PostInterview on how the average of results of multiple samples should be the expected value.

Expecting: It is useful in RL's case to recall some of his original responses on the PreSurvey because those responses help establish RL's initial contradictions in terms of his anticipation variation versus his occasional over-emphasis of the expected value. He demonstrated distributional reasoning even in the PreSurvey as he considered both centers and spread in his responses. In explaining his choices of "4, 5, 6, 6, 7, 8" for "Six Samples" of the Small Jar on PreSurvey Q1c, RL said that "while 6 red candies remains the average outcome, variation is likely." Later, when reasoning about 50 trials at the Small Jar on PreSurvey Q3, RL claimed that "a bell curve represents the most likely scenario – the extremes aren't seen often , the average is seen most often." As a final example, for "Six Samples" of the coin (PreSurvey Q7c), RL wrote that "while 25 flips are likely to be heads, in reality some variation is

likely, so my numbers represent a range that averages 25." RL's responses above show his consistency in combining the usage of both average and variation in his reasoning, even before the PreInterview.

However, RL was the *only* subject among the 27 who took the PreSurvey and put an unqualified "Yes" on Q1b when asked if results of several samples of the Small Jar would repeat. One other student put "Yes" but then qualified her answer, but RL alone was mathematically blunt. Six reds were expected on one handful because "reds are likely to be chosen according to their relative percentage of the total," and six reds would come out every time because "returning the candies recreates the original conditions, so the odds don't change." He reasoned similarly in comparing samples of fifty flips of a coin on PreSurvey Q7b, saying that "in the absence of any change of approach, the results [25 heads] are most likely to be the same." This latter response is very telling about the thinking of RL and how the expected value sometimes dominated his reasoning at the beginning of the research.

In the first few questions of the PreInterview, RL still had an occasionally unreasonable heavy emphasis on the expected value. For example, he thought he'd get "probably 6" reds for "One Sample" of the Small Jar on Pre Q1a, and he then reasoned that for "Several Samples" 6 reds would continue to be the most likely result to occur. RL liked list (iii) on PreInterview Q2, and indicated that though rare, he would not be "too surprised" to see all six samples result in 6 reds each.

It was in PreInterview Q9 and Q10 that RL seemed to really begin a shift in his expectations. RL initially put all tens for "One Sample" of sixty tosses of the die in

Pre Q9, saying that neither face of the die was "more likely or less likely than another." Because I knew that RL had a strong math background and was likely to simply rely on proportional reasoning, I spent extra time with RL to make sure he understood that the intent of Q9 was for him to put what really thought might happen if we did the experiment of tossing the die sixty times in Steve's class. At two different times, RL repeated: "I think that's going to happen [All tens]." Again, I found RL's attraction to the expected value was at times a powerful influence on his expectations. But in Q10, when asked to evaluate Lee's list of all tens and Lynn's narrow list, RL was quick to point out "I don't believe they actually got that." He started to defend why it was reasonable for someone to predict all tens, but then he reflected on his own earlier thinking about "Six Samples" of the Small Jar. He had listed a reasonable "4, 5, 6, 6, 7, 8" for his "Six Samples" on Pre Q1c, which was the exact same list he had put on the PreSurvey (Q1 was identical on both the PreSurvey and PreInterview). He explained that he had originally thought of putting all sixes in his list for "Six Samples" of the Small Jar, but that he knew six reds wasn't "exclusive to all other possibilities." Thus, when ruminating over his list of all tens on Pre Q9 and reflecting on Lee's list of all tens in Pre Q10, he said:

> RL: That [All tens] would be the basis of my expectations. But it would be pretty funny, to see the likelihood matched so closely. Just like the other one [Flips back to Q1c, "Six Samples" of the Small Jar], just like this one [Q9]. Well, this was like, when I originally did this [Q1c], and I put 6 down for the first one, and then I'm saying: You know what? We're living in the real world, this is not going to be 10, 10, 10...

He then changed his Q9 list to "5, 8, 9, 11, 12, 15", and I noticed that he didn't even

bother to include the expected value of 10 in his list. He was very articulate in explaining his change of mind:

> RL: I'm changing my mind [on Q9] because I'm making the same mistake that I was accusing some kids earlier of, and that I was considering average but not considering variation...You need to consider variation to get the full picture. This [All tens] is average only... there will always be a range of responses...but not every response will be 10. So, I think I was being limited in my consideration.

Pre Q9 and Q10 seemed to be watershed events for RL's thinking, in the sense that he really appeared to be engaged in some meta-cognition. He thought about how he had responded on earlier questions, and how repeated results of the expected value just didn't really make sense to him anymore. On PreInterview Q11, "Six Samples" of the die toss, RL listed "8, 9, 10, 10, 11, 12." He said that he didn't expect to see the same result for each of the six samples, and that he had "actually represented here a pretty limited range."

During the PostInterview, RL was clearly more appropriately attuned to range expectations than he had been in the PreSurvey or PreInterview. For "One Sample" of the Large Jar (Post Q1a), he initially said: "Well, I am inclined to estimate a range, rather than give an exact number." He then predicted "within a pretty wide range...I would say, even as low as oh, 40 to 80, even." After listing a reasonable "50, 55, 62, 65, 68, 70" for "Six Samples" of the Large Jar (Post Q1c), RL justified his choices by saying that "they are near the most likely value, but still wide enough to account for variation." Later, in PostInterview Q10a, RL thought that "One Sample" of the spinner would have a result "somewhere between 21 and 29...it's probably within that range." He felt in "Comparing Samples" for Post Q10b that results would "probably

not" be exactly the same each time, but that results were "likely to fall in a same range, similar range" as he had given for "One Sample" on Post Q10a. He also gave a reasonable list of "21, 22, 23, 27, 28, 29" for "Six Samples" of the spinner in Post Q10c, pointing out to me how "there's no repeats, but...they're similar, not identical." When I commented on how his list did not include the expected value, he said that "25 is the theoretical expected result, but that's not to say that that defines what happens."

RL's shift away from stressing the expected value went along with his marked decrease in expressing his mathematical calculations. While he stressed his mathematical computations in the PreInterview, he never gave voice to his calculations in the Post. In PreInterview Q1, RL wondered about the likelihood of getting all yellows (that is, no reds) in his sample of ten candies, and he calculated aloud:

> RL: Right. So, it is...if there is a 0.4 chance of pulling yellow, and then there's 0.4 chance of pulling another yellow, then there's a 0.16 chance of pulling two yellows. And if you've got ten, then you've got 0.4 to the tenth, which makes it real unlikely

Aside from the way that RL had considered drawing a candy and then replacing it (as opposed to the intent of the sampling scenario, which was to pull the handful without replacing any of the ten candies), his response was unique in that no one of the other cases showed such a willingness to calculate to the same extent as RL. He even speculated on a rate of convergence in PreInterview Q2, noting:

> RL: Because if you've got...since I'm talking about decimals, when you multiply them they get smaller and smaller. But at least when you're using the 0.6 [For getting Red] it gets smaller more slowly.

There were many other places with both PreSurvey and PreInterview where RL offered fairly intense mathematical computations to analyze the situation, but not after the die-tossing questions in PreInterview. In the PostInterview, he still flavored his explanations with relatively sophisticated mathematical terminology, such as when he wondered about finding the "inflection point" on a normal distribution of results, but he never articulated his calculations. Perhaps one reason for this shift is because his calculations were more useful to RL in finding centers, but not in estimating the variance that he came to expect.

RL's explanations in the PreSurvey and PreInterview often reflected distributional reasoning, but he was even more explicit about wanting a symmetric or skewed distribution in the PostInterview. For example, in PreInterview Q1a, when RL was talking about whether or not extremes were likely, he said:

> RL: Well, I mean that, it is entirely possible that if there were 99 candies and 1 yellow, you could pick that one yellow every single time. It is possible. It's on the far end of a bell curve, it's extremely unlikely, but it COULD happen.

I noticed how RL had used "bell curve" in his reasoning, even though there were no graphs provided for the question. It turned out that RL frequently envisioned what he thought the underlying distribution would look like, and occasionally he drew graphs on the interview script to help him make his point. In Pre Q1c, when predicting results for "Six Samples", RL started off saying "I'm going to use a bell curve, put 6 right at the middle here..." and then he drew a skewed bell curve and used it to help him make his choices. Later, in "Six Samples" of the die toss, RL mentioned that with

each sample of sixty tosses "the whole distribution would be different." Thus, I could see how the distribution of results was important to RL even in the PreInterview, and I noticed how his lists on the "Six Samples" questions were symmetric around the mean ("4, 5, 6, 6, 7, 8" on Pre Q1c; "5, 8, 9, 11, 12, 15" on the amended Pre Q9; "8, 9, 10, 10, 11, 12" on Pre Q11). During the PostInterview, he continued to refer to distribution, often making clear his preference for a symmetric or a skewed distribution. In "Comparing Lists" on Post Q2, he liked list (ii) "because the graph of this distribution skews to the left, [so] I would think I'd see more pulls higher than 60 than less than 60." After he gave his range of 21 to 29 blacks for "One Sample" of the spinner in Post Q10a, he explained that

> RL: Because there's gonna be a, uh, symmetrical distribution, neither of these is more likely than the other...which is why I don't say, you know, 18 to 29...I'm going afar from 25 in either direction.

RL therefore explicitly detailed how he wanted variation on both sides of the mean. In the "Six Samples" of the spinner, he justified his Post Q10c list of "21, 22, 23, 27, 28, 29" by saying that he "did a pair, each equally far out from the mean in either direction." He continued to stress distribution in "Comparing Lists" for the spinner on Post Q11. Although he said "I expect to see a symmetric distribution" in rejecting list (i), the symmetrical list (v) was also rejected as unlikely, because RL said:

> RL: Yeah there's variation, but there's so LITTLE variation, that it discounts the possibility that even though a wider distribution of values isn't AS likely, it is still SOMEWHAT likely, and so… I'm a little suspicious of such a tight distribution.

The reason he favored list (ii) for Post Q11 was "because there's a range that seems legitimately wide, and it looks at first blush to be relatively symmetrical."

Displaying: RL attended to all elements of the distribution (center, range, shape, and spread) when considering *displays* of variation in both interviews, but he misidentified real graphs as fake and vice versa in each interview. In PreInterview Q3, he thought "Graph: 30" was what he "would expect to see," even though "Graph: 30" was really made up. Although he also expressed some surprise at the limited range for "Graph: 30", he liked the fact that the mode was the expected value of 6 reds. He correctly thought "Graph: 300" showed actual results, and said that "the last graph here [Q3] was a very rough bell curve, [and] this [Q4] is much more similar to a bell curve" of the type he expected to see. For "Graph: 300", he noted that "you see other things that are POSSIBLE, just relatively unlikely. You still see them come up, but just less often." Thus, he commented on the shapes of both "Graph: 30" and "Graph: 300", and also the ranges. He offered similar reasoning strategies in the PostInterview. However, while he correctly identified "Graph: 30" as real in Post Q3, he thought that "Graph: 300" might be fake (when it was not) in Post Q4. His concern about "Graph: 300" in Post Q4 was that he said "I would expect to see more, more extreme values...It's kind of funny that there's nothing outside that certain range." I thought that in PreInterview Q3, at a time when RL was still fixated on the average, it was natural for him to accept the relatively tight range of "Graph:30" with minimal suspicion. He almost seemed to have over-compensated in his appreciation of range expectations in PostInterview Q4, since he thought that the range should be even wider than it already was.

A comparison of RL's responses to the "Likelier Graphs?" questions in both interviews showcases his reasoning skills in evaluating and comparing graphs while also demonstrating an improvement in his final conclusions. In PreInterview Q13, RL thought that Group A was the fake graph and that Group B was the real graph, when the opposite was true. About Group A, he said:

> RL: [Pre Q13] You have such a wide range...more outliers in Group A....Not only did someone get the very unlikely result of 13, somebody else got the very unlikely result of 30...Both of those [extreme values] are further out than everything else.

I noticed that RL never commented on how Group A only had one out of twenty results at the expected value of 20 blacks, but that his explanation hinged upon the wide range of Group A. His reasoning on Group B included a reference to center and spread:

> RL: Group B is tighter, and definitely holds to a center more...in Group B, you just didn't see that kind of variance [as in Group A]. In Group B, people did it, and no big surprises.

It did not surprise me that RL erroneously thought Group B was authentic, because the graph for Group B was almost symmetric about the mean, just as all of RL's lists were when he predicted results for multiple samples. In PostInterview Q13, RL correctly identified Class A as fake and Class B as real. He suspected that Class A had "underestimated, perhaps, the possibility of seeing less common values," meaning that he thought there should be more extremes in the graph for Class A. Furthermore, he said "the shape here [for Class A was] very tight, very narrow, not a lot of variation", and he thought the data was unnaturally grouped around the expected value of 25

blacks. Class A was "too neat" for RL to believe it came from genuine data. In Class B, on the other hand, RL saw "a lot more different values being represented," which he liked. He said he would have expected to see a graph like Class B's coming from real data because:

> RL: [Post Q13] You still see in the middle there, between 24 and 27 or what have you, you see most values. And then a few on either side, kind of trickling out, or sprinkled to the sides.

I appreciated how RL attended to spread, noting that he pointed out "most values" in a central subrange of 24 to 27, and yet he also was comfortable with the few extremes shown in Class B.

RL used similar reasoning when comparing distributions on both the "MAX Wait-Times" question in the PreInterview and the "Muffin Weights" question on the PostInterview, but he increased his attention to subranges in the latter question. For PreInterview Q8, he thought the Eastbound train was more "reliable," and he pointed out how the Eastbound train had a shorter range than the Westbound. For the Westbound train, he made a point of stressing how a potential passenger wouldn't know what to expect for wait-times, because "there's a lot of variation...it's not a consistent pattern." When I asked him to comment on someone's argument that there was no difference in wait-times because the averages were the same, he said "the average does not tell all the story...they are not including the variation." He then affirmed his initial view that "the Eastbound is more predictable, and less variation."

For the "Muffin Weights" of PostInterview Q9, RL named the West End bakery as being "a little more reliable." When I asked him why, he said that "the interquartile range is narrower, you can pretty much count on most muffins are gonna be within a certain range." I thought that RL used the boxplots in Post Q9 appropriately, and he went on to comment negatively about the wider range of the East End bakery. Just as he had wondered about how long a passenger might wait for the Westbound train in the PreInterview, so too RL was concerned about the weight of the muffin he might get at the East End bakery:

> RL: The variation is SO much that, if I'm looking for reliability, if I wanna know what I'm gonna [get], to expect, then I don't wanna mess around wondering if I'm gonna get a huge honkin' muffin, or I'm gonna get a little sub-standard muffin...

He had a good observation in comparing the relative merits of boxplots versus dotplots, noting that "with the boxplot you're sacrificing information." When he wanted to see the specifics of where the data was grouped, he found the dotplots more helpful.

Interpreting: RL referred many times in the PreSurvey and PreInterview (but not as much in the PostInterview) to the way that theoretical probability was different from reality, and he seemed to use this theme to convince himself that results would not be the expected value each time. For example, in PreInterview Q1c RL had explained to me why he had initially put all 6s on the identical question earlier on the PreSurvey Q1c:

> RL: The first thing I did for part c [PreSurvey Q1c], where how many do you [think you'll get], and I wrote "6" in every one.
> I: Oh, yeah.

RL: Being very strict as in probability-dictated reality, as distinct from described likelihoods. And so I went back and, instead of 6 every one, this one 6 and then a 5 and a 7, and a 3...
I: So you changed it [on the PreSurvey] ?
RL: I did change it, when I went back and I thought, okay, reality is going to impinge on the strict likelihood by a given thing

Later, in PreInterview Q3, even though RL thought "Graph: 30" was authentic, he wondered if one might see more extremes "just because weird things happen" in real life. Again, he used "living in the real world" as a reason for changing his list of all tens in "One Sample" of the die toss in PreInterview Q9, mentioning our "world of imperfect scientific conditions." He stressed that real-world "conditions are never identical", and that was a reason he thought he would see variation in results. On the PreSurvey Q7b RL had implied that absent "any change of approach" in flipping coins, results would "most likely...be the same," and yet by the end of the PreInterview he seemed to believe that variation was unavoidable. He did not say as much about the difference between reality versus theory in the PostInterview, perhaps because he had already convinced himself of the difference in the PreInterview and throughout the class interventions.

RL consistently described influences of the number of samples in both interviews, and he included many explicit references to the influence of doing *fewer* as well as *greater* numbers of samples. One characteristic that he frequently pointed out was how the number of samples affected the chances of getting extreme values. For example, in PreInterview Q4 RL had envisioned his own hypothetical jar containing 99 Red candies and 1 Yellow candy:

RL: Just like the example I gave with the one yellow candy and the 99. You can pull out the yellow candy, it's possible. It's not going to happen that often, but if you do it enough times, sooner or later, it's bound to happen.

RL was even more specific about the number of samples in PreInterview Q11, when he considered "Six Samples" of the die toss. He had listed "8, 9, 10, 10, 11, 12", which we both thought had a "pretty limited range." RL justified his choice as follows:

RL: But we're only talking about 6 people throwing, and when you've got 6, it's a pretty small sample size. So, chances are you're not going to see anything too goofy. You get a hundred people doing this, you're definitely going to see the extremes pop up more often.

I asked him what kind of range he might expect with one hundred samples, and he suggested the maximal range possible (0 to 60), saying "It can happen. It is unlikely." I thought RL had an under-appreciation of just how unlikely it is to get sixty 5s in sixty tosses of the fair die, but after the class interventions he had a better sense of how many samples it might take to attain extreme results. In considering samples from the Large Jar in PostInterview Q4, he talked about the chances of getting 0 red in his sample of 100:

RL: It's possible, it's a one in..., you know, trillion badillion, but if you do a hundred bunillion trillion pulls, you're gonna get a zero! And so, the more pulls you do, the more opportunity that exceptional event has of occurring.

He expressed the idea of more samples offering more chances to attain extreme results at many other times in the PostInterview, and was suspicious of unexpected results occurring with few samples. For example, in Post Q11 he declared list (i) for six

samples of the spinner ("38, 43, 36, 26, 41, 33") to be high overall, and then he added:

> RL: So, I'm a little bit suspicious, that having done so few spins, there's so many relatively unlikely [results]…With more sets, sure, I think again you're gonna start to see these more exceptional things happen, but you will also have seen many more expected results.

The last part of RL's previous response illustrates how RL also thought that more samples gave more chances to actually attain the expected value. In fact, he generally thought in both interviews that the average results of multiple samples should be the expected value. In other words, he did not expect the mean, median, or mode to vary when doing more samples. Even in PreInterview Q9, when he finally talked about "living in the real world" and how results for sixty tosses of the die would not be all tens, he stressed that "if you're going to see a range, the average of that range will be 10." During the PostInterview, he repeatedly made it clear that results from multiple samples should have an average equal to the expected value. For example, in "One Sample" of Large Jar (Post Q1a), he thought that "over time, if I pulled 100 candies, put them back, pulled another hundred candies… I think I would average a representative of 60 red , 40 yellow." Again, in "Several Samples" of the Large Jar (Post Q1b), RL thought that he'd "get more 60s than anything else", suggesting that the mode should be the expected value. He was even more explicit in PostInterview Q3 when he expressed how credible "Graph: 300" was, saying "we've got a mode of 60, which is what I would expect to see, so it looks believable."

Finally, in both interviews RL emphasized the influence of the number of samples on the shape of distribution of results, and he did so to a much greater extent

than any other subject. During the PreInterview, for instance, as he compared "Graph: 30" and "Graph: 300" toward the end of his analysis in Q4, he noted that:

> RL: I expect to see a certain bell curve, given more trials. This was to so few trials [in "Graph: 30"], that it's not a very fleshed-out bell curve. Here [in "Graph: 300"] you start to see things fall into a pattern. If you did this [sampling] ten thousand [times], you'd probably have a really nice bell curve. So, I attribute the more bell curve-looking design to the number of trials.

I was impressed at RL's articulation in connecting the shape of the distribution to the number of samples, and he expounded on that connection even more frequently in the PostInterview. Even when no graph was present, as in "Several Samples" of the Large Jar in PostInterview Q1b, RL's language reflected the shape of the underlying distribution:

> RL: [Post Q1b] Well, the more times I draw, the more normal the distribution, I think I'd get more 60s than anything else, but the more you draw, then the wider the distribution as well. More, just – The more you draw, the more chance there is of getting an outlier, or an extreme value. So, I would think that the more I draw, I'm more likely to get… Well, over time I think I'm more likely to get within a tighter range, actually.

At first I thought RL had concluded that results of a greater number of trials would have a smaller overall range than would the results of a fewer number of trials. However, based on all his other responses, particularly those having to do with distribution, it seems more likely that what RL meant was that data for more trials would likely be more concentrated within a narrower subrange. His last comment in the PostInterview was a good exemplar of his view about the influence of more

samples on the shape of distributions. After he considered the "Likelier Graph?" in Post Q13, he said in conclusion that "it's easy to see how more sets will start to normalize that distribution and approach the theoretical prediction."

In conclusion, RL had the broadest and strongest mathematical background of anyone else in the MET 2 class. He demonstrated distributional reasoning even in the PreSurvey as he considered both centers and spread in his responses. However, RL's thinking was occasionally over-influenced by the expected value in the PreSurvey and much of the PreInterview. After the die-tossing questions of the PreInterview, RL more consistently expressed his expectations in terms of ranges. Although he never once referred to what he had seen in class, I think that the class experiences and his own self-reflection led to an improved sense of variation. His mathematical computations, which he articulated in the PreSurvey and PreInterview but not in the PostInterview, seemed to influence his choices throughout the research. Thus, RL carefully predicted results for multiple samples so that they were symmetrically distributed about the expected value. He reasoned about graphs using all aspects of the distribution (center, range, shape, and spread), and seemed to pay even more attention to relative subranges on the PostInterview. However, RL incorrectly identified real versus fake graphs on both interviews. Although he commented several times before the class interventions how reality was different from theory, after the interventions he stopped making those kinds of comments. Lastly, RL initially had a reasonable sense of the influence of the number of the number of samples on the distribution of results, and he demonstrated an even more extensive understanding of this theme in the PostInterview.

Cross-Case Comparisons

As mentioned previously, responses can be coded at multiple places within the framework, a possibility that arises when a response is longer and multi-faceted. From the interview transcripts, I took questions or portions of a question and considered each case's response through the evolving framework. Because some questions had multiple parts, there were often some substantial and lengthy responses by a case to a question. To illustrate what I did, consider Q2 ("Compare Lists" for the Small Jar) on the PreInterview. I asked subjects to pick the list(s) that they thought might be likely to occur as choices for six trials, and then to comment on all the lists. Then I asked them which list they thought *best* described what might happen and explain why. Since there were five lists, naturally this interview question had the potential to elicit a considerable amount of dialogue in response.

In the cross-case analysis, I took each question or subquestion (such as Q1a, Q1b, Q1c, or Q2), and coded the aggregate response for each case. That is, I took everything the subject said on that question or subquestion and saw how the parts of the response fit into the framework. On Q2, for instance, I generated Table 13 to show me how the different cases' responses fit into the framework. I called such tables "CodeFrames" because I was coding responses in view of the framework. I made CodeFrames for every question on the PreInterview and PostInterview, including subquestions as I thought necessary or advantageous.

Table 13. *CodeFrame for PreInterview Q2 (Cross-Case Analysis)*

Framework	Description Within Themes	Subject (Case)					
1Ai	Should be on Both Sides of Exp. Val.	DS	EM		JM	RL	SP
1Aii	Won't be Exp. Val. Each Time	DS	EM	GP	JM		SP
1Aii	Shouldn't Repeat Values in General						SP
1Aiii	Should be in the MidRange		EM	GP			SP
1Aiii	Shouldn't be Too Many Highs (or Lows)		EM				SP
1Aiii	Should be Within Range Around Exp. Val.		EM		JM		SP
1Bi	Expected Value is Most Likely	DS			JM	RL	
1Bi	Extremes are Unlikely	DS	EM	GP	JM	RL	
1Bi	Extremes are Possible	DS				RL	SP
1Biii	Proportional Reasoning				JM	RL	
3Bii	Nature of the Candy Mixing		EM				
3Ci	Anything Can Happen				JM		
3Cii	Difficulty in Making a Choice			GP			
3Dii	Expected Value as an Average		EM				
3Dii	More Trials = More Variation	DS					

The CodeFrames give much information: the rows give different themes from the framework, or specific characteristics within the themes. The columns under the Subject (Case) heading show which cases were coded at the different places within the framework. For example, in Table 13, the CodeFrame for Q2 on the PreInterview shows how DS responded throughout Q2. Reading all the way down the Subject column for DS, we see that she had some part of her response address how more trials would give more variation. Moving across the row for "More Trials = More Variation", we see that no one else but DS included that theme as part of a response for Q2 on the PreInterview. On the other hand, Table 13 shows how five of the six cases all addressed three characteristics of themes in their responses to Q2. Results should be close to the expected value [1Ai], results won't be the expected value each time [1Ai], and extremes are unlikely [1Bi]. In this section I will

summarize the rows (which represent dimensions, themes, or characteristics of themes from the framework) from PreInterview and PostInterview questions where all six cases had part of their response coded at that row.

As mentioned above, I had many CodeFrames for each of the interviews. The number of rows in each CodeFrame was inherently variable, depending on what the cases had to say. For instance, Table 13 has fifteen rows simply because that's how many dimensions, themes, or characteristics of themes occurred in the collective responses of the six cases. In some of the CodeFrames, there are matches among all six cases for certain rows, which I refer to as "Match 6 Rows". There were also "Match 5 Rows", meaning that exactly five of the six cases were coded along that row (there are three such rows in Table 13). Table 14 shows the number of CodeFrames for the questions in each interview, the number of rows in each CodeFrame, and how many Match 5 or Match 6 Rows were in each CodeFrame.

Table 14. *CodeFrame Summary*

PreInterview				PostInterview			
Question Number	Rows in CodeFrame	Match 5 Rows	Match 6 Rows	Question Number	Rows in CodeFrame	Match 5 Rows	Match 6 Rows
Q1a	12	2		Q1a	11		1
Q1b	10			Q1b	10	1	
Q1c	11			Q1c	7	1	
Q2	17	3		Q2	14	2	3
Q3	10		1	Q3	12	2	
Q4a	12	1	1	Q4a	12		
Q4b	6			Q4b	8		
Q5	16		1	Q5	16		
Q6	4	2		Q6	6	2	1
Q7	18	2	2	Q7	17	3	1
Q8	15		2	Q8	13	3	1
Q9	7		1	Q9	16	1	
Q10	17	3	3	Q10a	8		2
Q11a	9	1	1	Q10b	6	1	1
Q11b	9		1	Q10c	10		1
Q12a	4		1	Q11	15	5	3
Q12b	6	1		Q12	14	1	3
Q12c	6			Q13ab	10	1	
Q13a	7			Q13c	14	1	1
Q13b	17						
20 Total	213 Total	15 Total	14 Total	19 Total	219 Total	24 Total	18 Total

I chose to show Match 5 or Match 6 Rows in Table 14 because those were the strongest two levels of agreement. As a percentage, the Match 5 or Match 6 Rows out of the total rows are [(15+14) / 213] = 13.6 % for the PreInterview and [(24 + 18) / 219] = 19.2 % for the PostInterview, suggesting slightly more agreement in the PostInterview. The fourteen Match 6 Rows in the PreInterview and the eighteen Match 6 Rows in the PostInterview are summarized next, because they represent the most agreement.

PreInterview

Table 14 does not indicate the parts of the framework which garnered agreement, unless we see the actual rows from the different CodeFrames. Thus, I cut the rows out of each of the relevant CodeFrames and used the framework to re-organize the fourteen Match 6 Rows from the PreInterview. Table 15 shows exactly what the fourteen rows represented with respect to the framework.

Table 15. *Match 6 Rows from PreInterview CodeFrames*

Framework	Description Within Theme or Dimension	Question
1A	Riki: Really rolled It	Q10
1A	Yes, I'd be surprised if more Black than White in 3 spins	Q12a
1Aii	Repeated values could happen	Q11
1Aii	Their own choices are all different	Q11
1Bii	Probability arguments (chance, likelihoods)	Q9
1Bii	Extremes possible, but unlikely	Q10
1Biv	Lynn: Not enough variation	Q10
2Bi	Focus on mode in comparing graphs	Q7
2Bi	Comments on same summary statistics	Q8
2Bii	Noticing limited range: Only got three types of values	Q3
2C	Those are the real results shown in the graph	Q4
2C	Comfortable with average answer for "True duration of trip"	Q7
2Ci	Eastbound train: More consistent or reliable	Q8
2Cii	Engineer: Should use Graph 3	Q5

Table 15 shows where within the framework there was agreement among all six cases on the PreInterview. I'll discuss some of the areas of agreement from Table 15 in terms of the evolving framework.

Within the aspect of *expecting* variation, notice how all six cases thought Riki was the student on Q10 ("Who Cheated?") who really rolled the die. There is no theme within the dimension for that row, meaning that it is simply in the framework as

[1A] because it shows specific expectations for that question. Most importantly, Riki did have the list which showed genuine data, and the fact that all six cases correctly identified Riki as having really rolled the die shows reasonable expectations on the part of the subjects. The theme concerning repeated values [1Aii] also had agreement in Q11 ("Six Samples" of sixty tosses of the die), and every one of the cases gave a list of predictions that contained distinct values. In other words, the lists given had no repeated values. I think that the subjects were particularly careful to choose all distinct values for their list of "Six Samples" in Q11 because the discussions of Q9 and Q10 tended to bring out strong reactions from the subjects about how six results of sixty would not occur. Thus, the subjects may have been over-compensating, thinking that if six results would not all be identical, then the six results would be all distinct. In explaining *why* they held their opinions, the cases agreed in Q10 ("Who Cheated?") that extremes were possible but unlikely [1Bii], and that Lynn's list did not exhibit enough variation [1Biv]. I was surprised that all the cases were suspicious of Lynn's list, because I had thought some subjects might argue that "Anything Can Happen." However, the common sense from the cases was that Lynn's was too narrow, which is a reasonable assessment.

In *displaying* variation, all cases focused on the averages [2Bi] shown for the graphs in Q7 ("Compare Graphs") and Q8 ("MAX Wait-Times"). Since averages are such a dominant part of traditional stochastics curricula and the media, it was no surprise that my subjects' attention gravitated towards centers. I was encouraged to see that all subjects commented on the narrow range depicted in the graph on Q3 ("Graph: 30"), which did show fabricated data. However, even though they had a

focus on the range [2Bii], not all the subjects identified the graph as being fake. In Q4 ("Graph: 300") the same specific conclusion [2C] for that question was made by all cases: Those were in fact real results shown in the graph. I also noticed that when making conclusions about the Eastbound train in Q8 ("MAX Wait-Times"), all cases had some emphasis on the consistency or reliability [2Ci] of the train. Finally, there was agreement that the Engineer of Q5 ("Car Brakes") should use Graph 3 in her report, a reasonable conclusion to make in the context of the question [2Ciii].

PostInterview

When I realized that there was agreement (Match 6 Rows) in the PreInterview for the *expecting* and *displaying* aspects but not for the *interpreting* aspect, I was curious to see how the eighteen Match 6 Rows for the PostInterview were organized according to the framework, and this organization is given in Table 16. Not only was there at least some agreement in the PostInterview on all three aspects, but the nature of the agreement represented an overall maturity of reasoning about variation. I'll comment more on this observation after discussing some of the areas of agreement on the PostInterview in terms of the framework.

Most of the agreement in PostInterview responses had to do with *expecting* variation. For example, on Q1a and Q10a ("One Sample" of the Large Jar and spinner, respectively) each of the cases did not just put the expected value for their prediction, but rather they gave answers indicating an appreciation for variation.

Table 16. *Match 6 Rows from PostInterview CodeFrames*

Framework	Description Within Theme or Dimension	Question
1A	Gives a # Other than 60 or Range	Q1a
1A	Gives a # Other than 25 or a Range	Q10a
1A	Picks List (ii)	Q11
1Ai	Should be Close to the Expected Value	Q11
1Ai	Should be Close to the Expected Value	Q10b
1Aii	Their own choices are all different	Q10c
1Aii	Shouldn't repeat: Should be Different	Q2
1Aiii	Should have Variation or Range	Q2
1Bi	Extremes Unlikely	Q2
1Bi	Extremes Unlikely	Q11
1Bi	Extremes Possible	Q12
1Bi	No Guarantee of Getting Expected Value	Q12
1Biii	Proportional Reasoning	Q10a
2C	Class A : Likely Cheated	Q13
2Ciii	Rounding Affects Accuracy	Q7
2Ciii	More Detail in Histogram	Q8
3Bi	Operator Method or Perspective in Using the Scale	Q6
3Dii	Number of Spins Affects Amount of Variation	Q12

The six cases also all favored list (ii) on Q2 ("Compare Lists" for the Large Jar), which was the most reasonable choice. All cases gave responses that reflected the theme concerning the expected value [1Ai] in Q11 ("Compare Lists" for the spinner) and in Q10b ("Compare Samples" for the spinner). In particular, the cases' responses indicated that results should be close to the expected value. Further agreement for *what* was expected included the themes concerning repeated values [1Aii] and the idea that results should exhibit a range or some variation [1Aiii]. Regarding reasons *why* expectations were held, the language of possibilities and likelihoods [1Bi] was used by all cases in response to several questions: Q2, Q11, and Q12 ("Compare Comments"). One key idea that seems commonly held is the notion that extreme values are possible but unlikely.

For *displaying* variation, everyone commented on Q13 ("Likelier Graph?") that the graph for Class A was likelier to reflect made-up data, a correct conclusion [2C]. Concerning level of detail and usefulness of different types of graphs [2Ciii], there was consensus that less rounding led to a more accurate graph in Q7 ("Compare Graphs" for the muffins) and also that the histogram showed more detail in Q8 ("35 Muffins"). I noticed that there was no agreement for specific characteristics of themes within the dimension of evaluating and comparing graphs for the PostInterview, and I suspect one reason is that the questions offered more graph types than on the PreInterview, hence more opportunities emphasize themes in different ways.

There were two dimensions of agreement in *interpreting* variation. One dimension concerned causes of variation, in that all cases had some theme of operator error in using the scale for the repeated-measurement question involving the weight of a single muffin (Q6: "Causes: Muffin"). I considered the causes they listed as naturally occurring causes [3Bi] because they did not include a deliberate, subversive attempt to introduce variation, but were the kinds of variation that one would reasonably expect to find among different people attempting to discern a measurement. Finally, in Q12 ("Compare Comments"), everyone had some element of their response that connected the number of trials or spins with the resulting variation [3Dii].

In summary, the use of the CodeFrames in the cross-case analysis reveals some overall trends, most notably the closer agreement in *expecting* variation in the PostInterview. For example, on the PreInterview there were many predictions for "One Sample" questions that were just the expected value. However, on the

PostInterview ranges were given for predictions, or values that were explained as being "near" to the expected value. Also, on the PreInterview, some cases did not access proportional reasoning, while others seemed overwhelmingly influenced by theoretical predictions. On the PostInterview they all used proportional reasoning but no one claimed that the theoretically expected value should always be the outcome. There were also some uniformly reasonable conclusions made regarding *displays* of variation, as well as attention given to average and range when evaluating and comparing graphs. Finally, in the PostInterview there was total agreement about plausible *interpretations* of variation, namely the cause of variation in Q6 ("Causes: Muffin") and the influence of more trials on results in Q12 ("Compare Comments").

CHAPTER FIVE

Discussion and Conclusion

This chapter summarizes the main contributions of my research findings to the general field of studying probability and statistics education among teachers and students. I'll discuss how the study has addressed each of my research questions in turn, offering further analysis and reflection concerning teacher education before articulating some limitations of the research. Then I'll outline some implications for future research.

First Research Question

The first research question asked "What are the components of a conceptual framework that help characterize EPST's thinking about variation?" The evolving framework presented in the first part of Chapter 4 informs this question by offering an in-depth exploration of a sample of elementary preservice teachers' thinking about variability. The framework, reproduced in Figure 32, provides a lens through which three different *aspects* of an EPST's understanding of variation can be viewed. The three aspects address how EPSTs reasons in terms of *expecting*, *displaying*, and *interpreting* variation.

Expecting

When *expecting* variation, my subjects expressed both *what* they expected and *why*. The expected value or average was a frequent theme concerning *what* EPSTs thought might occur. A dominant type of response was how results should be close to,

about, or near the expected value, and a more explicit type of response was how results might be higher or lower than the expected value.

Evolving Framework

[1] Expecting Variation
- A] Describing What is Expected
  - i) Concerning Expected Value
  - ii) Concerning Repeated Values
  - iii) Concerning Range or Extremes
- B] Describing Why (Reasons for Expectations)
  - i) Involves Possibility or Likelihood
  - ii) Involves Experiential Reasoning
  - iii) Involves Proportional Reasoning
  - iv) Involves Distributional Reasoning

[2] Displaying Variation
- A] Producing Graphs
  - i) Technical Details
  - ii) Characteristics of the Distribution
- B] Evaluating & Comparing Graphs
  - i) Focus on Average
  - ii) Focus on Range or Extremes
  - iii) Focus on Shape
  - iv) Focus on Spread
- C] Making Conclusions about Graphs
  - i) Emphasizing Decisions in Context
  - ii) Emphasizing Consistency or Reliability
  - iii) Emphasizing Level of Detail & Usefulness

[3] Interpreting Variation
- A] Defining Variation
  - i) Definitions & Descriptions
  - ii) Examples
- B] Causes of Variation
  - i) Naturally Occurring Causes
  - ii) Physically Induced Causes
- C] Effects of Variation
  - i) Effects on Perception
  - ii) Effects on Decisions
- D] Influencing Expectations and Variation
  - i) Quantities in Sampling
  - ii) Number of Samples

Figure 32 – Evolving Framework

Responses of both types show an acknowledgement of variation, and offer the teacher educator a good stepping stone to finding out just how close or how much higher is reasonable. Another theme for *what* was expected concerned whether or not results from multiple samples might repeat. Almost all of my subjects thought that results would not necessarily repeat each time, but there were a few who either implied or stated outright that results should or would repeat every time. It is crucial for a teacher educator not to assume that EPSTs automatically believe results are unlikely to repeat each time. An EPST may have had little or no exposure to probability and statistics and may actually believe that point estimates offer a "correct" answer to questions involving multiple samples. For example, the thinking may be that if "6 reds" is a good guess for a single sample of the Small Jar, then "6, 6, 6, 6, 6, 6" is a good guess for six samples. A third theme for *what* was expected concerned ranges or extreme values. More specific than just suggesting that results be above and below the expected value, some responses actually specified a numeric range. When a person volunteers an answer like "between 21 to 29 black" for the result of one sample of fifty spins of the half-black and half-white spinner, that kind of answer shows the specific variation the person expects.

In describing *why* they held their expectations, almost all of my subjects' reasoning at some point involved the language of possibilities and likelihood. For example, many subjects explained how extreme results were possible but unlikely. In the absence of quantitative indicators, informal language (such as "entirely possible" or "highly unlikely") provides at least some qualitative gauge of just how possible or likely people perceive events to be. Students using informal notions of possibilities

and likelihoods can be challenged to come up with their own quantitative measures of how likely they think events are. The students should then be provided experiences by which they can test their ideas. Reasons involving experience constituted another theme for *why* my subjects held their expectations. Some subjects mentioned informal out-of-class experiences, such as games they had played, and other subjects recalled their experiences with the in-class activities. Experience can be very useful if subjects have a good way to record what actually happened, otherwise their recollection of events may be incomplete. For example, someone may claim to recall "hardly ever getting of ones" when tossing a fair die, when in reality their data may support a reasonable amount of ones. A person who thinks extreme results are "pretty likely" because they recall getting a sample result of 23 reds from the Large Jar may not remember how many total samples they had taken before obtaining that rare result.

The theme of proportional reasoning can be a useful anchor to help center expectations appropriately, and this theme was a part of many of my subjects' reasons for *why* they expected what they did. An over-reliance on proportional reasoning can lead to a restricted expectation of variation, but an under-reliance on proportional reasoning can also lead to poor expectations. For example, when SP expected samples from the Small Jar to be near her "midrange" results of 4, 5, or 6 reds, it was encouraging to note her expectation of variation but it was discouraging to see her that her choices weren't centered around 6 reds. In explaining *why* certain results are expected, the theme of distributional reasoning focused on elements of the distribution of data: Center, range, shape, and spread. Responses for this theme generally reflected a more comprehensive sense of variation than the other themes for the dimension of

describing *why* expectations are held. Even brief answers could reflect distributional reasoning, as in JL's response to Probability PostSurvey Q1b. The question asked how a second sample of 50 spins of the half-black and half-white spinner would compare to the first sample, and JL said "I think the sample results would get tighter, the grouping would accumulate around 50 [blacks]." She misstated the center as 50 blacks when it should be 25 blacks, and her answer also went beyond the idea of only two samples to anticipate multiple samples. However, the larger point is that JL conveyed a sense of the underlying distribution in her reasoning, and how she saw the data getting spread about the mean.

Displaying

Concerning displays of variation, EPST's showed their skills and reasoning along the three dimensions of *producing graphs*, *evaluating and comparing graphs*, and *making conclusions about graphs*.

When I considered how my subjects *produced graphs*, I could tell that the technical details of their graphs were a reflection of the subjects' graph sense. For example, I saw many smooth bell curves that were drawn for PreSurvey Q4, even though a bar chart or dotplot would have been a better choice for the type of graph in that problem situation. Some graphs had detailed axes with an appropriate scale, while some others had unlabeled axes or inappropriate scales. It is hard to convey a proper sense of variation in a graph if the student lacks a useful graph sense to begin with. Again, it is important for the teacher educator to not take for granted a student's graph sense, but to provide plenty of opportunities to assess and develop graph skills.

How the characteristics of the distribution get conveyed is another theme in *producing graphs*. My subjects generally gave centers that were reasonably placed, but they often provided ranges that were too wide. Spreads were occasionally too tight or too scattered, and shapes were often unnaturally symmetric. I suggest that teacher educators have their students gather actual data, make the resultant graphs, and discuss the specific ways that the graphs reflect characteristics of the distribution. As more types of graphs are learned, more possibilities are created for comparing how different graph types lend themselves to displaying variation.

When *evaluating and comparing graphs,* the four themes I looked for in my subjects' responses corresponded to the four components of distributional reasoning: Average, range, shape, and spread. The first of these themes, a focus on average, was reflected in most but not all of my subjects' responses. Many subjects were able to move beyond a focus on average to include references to other features of the distribution, but some people made it clear that the average was their primary consideration in answering any question having to do with graphs. The theme focusing on range or extremes was often reflected in questions having to do with which graph had more variation. In fact, many subjects seemed to make an initial association between "more variation" and "wider range" (and vice versa), which is not a bad first step to make in thinking about displays of variation. Not many subjects volunteered written information to indicate that they were attending to shape as a theme, but there were more responses that focused on shape during the in-class discussions and during the PostInterviews. My subjects had some standard ways of talking about shape, using language like "symmetric" or "skewed", "normal" or "uniform". There were also

some non-standard ways of referring to the shape of a distribution, including the use of hand gestures to try to communicate the picture in the subjects' mind. Responses that focused on the theme of spread depended on the type of display that subjects were considering. For example, when using dotplots some subjects referred to the way data was "clustered" or "scattered" along different places along the horizontal axis to indicate how they saw the way data was grouped or spread out in a graph. In using boxplots, we had discussed as a class how the interquartile range was one measure of spread, and many subjects referred to that measure in their responses.

The first theme I had for *making conclusions about graphs* was how some responses emphasized the context of the problem. For example, many subjects considered their own preference for rain, their tolerance in waiting for trains, and their desire for weighty muffins in making their conclusions. The context of the problem is particularly important for tasks involving graphs because the context invites the subject to consider whether or not the amount of variation shown is desirable or appropriate. In other words, it is not enough to merely analyze the graph, but instead it is desirable to think about what the variation means in the context of the problem. Responses in the next theme emphasized consistency or reliability of the phenomena under consideration. Subjects often associated a wider range with less consistency, and a narrower range with more consistency. Another theme for responses emphasized the level of detail and subsequent usefulness of the graph. This latter theme was particularly relevant in the questions which offered different types of graphs for the same data set. Some subjects clearly expressed their preferences for one type of graph over another, saying for instance how boxplots gave a good overall

sense of the data while histograms were too finely detailed. Because different graphs convey different messages about variation, it seems important to encourage student discussion about how useful a graph is.

Interpreting

The four dimensions that arose in the data for this aspect were *defining variation*, discerning *causes of variation*, *effects of variation*, and also *influencing expectations and variation*.

The two themes for *defining variation* were an actual definition of variation and also examples of variation. When I asked on the PreSurvey what variation meant to my subjects, many of them of them conveyed the idea that variation meant having differences, or the degree of difference. After all the survey and interview data had been gathered, I looked back on all the ways in which variation was reflected in the collective responses in order to further describe what students were saying about "variation." Two important uses of the term were to describe the range of data and to describe the distribution of data within the range. Other descriptors I found for variation included the way that data was clustered, spread, concentrated, and distributed.

The theme of examples of variation was addressed in the PreSurvey, and the main types of examples reflected natural or personal characteristics. Throughout the subsequent research instruments, students affirmed examples of variation via their responses. For instance, students provided choices on "Six Samples" that varied, and they gave graphs illustrating expected variation. Furthermore, they talked about the variation they did or did not expect in all contexts of sampling, data and graphs, and

probability situations. Thus, the students acknowledged that situations such as daily rainfall, muffin weights, MAX wait-times, or the results of a probability or sampling experiment are examples of things that vary.

Under *causes of variation*, one theme I saw reflected in the data was naturally occurring causes. My subjects had no trouble coming up with many different naturally occurring causes of variation, such as all the causes they listed for differences in rainfall patterns. In the sampling and probability situations, many students seemed to point to randomness as a naturally occurring cause. For example, if asked why the spinner doesn't land in the same spot each time it is spun and a student says that "it's luck," that may be the student's way of identifying randomness as the underlying cause of variation. The other theme of physically induced causes included those causes which were deliberate or intentional as opposed to naturally occurring. For example, lining up the spinner in the same spot for each spin and trying to apply the same amount of force each time was seen a physically induced cause for reduced variation. Class discussions about the different types of causes of variation can lay a good foundation for the notion of variation which we can and cannot control, which in turn can help students generate ideas about how to minimize variation.

I thought of *effects of variation* in terms of two distinct but related themes, the effect of variation on students' perceptions and the effect of variation on their decisions. For the first of these themes, some students perceived a difference between theoretical predictions and real-life outcomes. Also, many students perceived that "Anything Can Happen" in situations involving variation. The second theme

concerned the effect of variation on students' decisions. Some students expressed a lack of confidence in making decisions, and "I Don't Know" was a frequent response reflecting this theme. In making inferences, it seems that the two themes for *effects of variation* were often linked. For example, a student who thinks that "Anything Can Happen" may therefore think that there is no way to decide what might happen, and thus the student may respond with "I Don't Know".

The two themes for *influencing expectations and variation* were quantities in sampling (i.e., the numbers of candies in the population or in the sample) and also the numbers of samples. The first theme applied primarily to the context of drawing samples where there was a discrete population, such as samples of candies from the Small Jar or the Large Jar. Several subjects focused on the sheer numbers of candies in the jar, and in some cases it seemed that the probabilities of getting different outcomes were linked to quantities. Particularly for subjects who are not strong proportional reasoners, there may be a tendency to see the quantity and not the ratio as the influential factor in what the sample results are likely to be. For example, if getting a sample result of 9 red is unlikely for the Small Jar, then a student may reason that a sample result of 90 red is much more unlikely for the Large Jar since there are "so many more candies".

The second theme, involving the numbers of samples, was reflected in many different ways. Almost all of my subjects pointed out that more samples would widen the overall range, while very few subjects suggested that more samples would also tighten the subrange capturing most of the results. Other ideas included how more

samples offered more chances to attain the expected value, and how more samples provided a better picture of the underlying distribution.

The summary of the evolving framework provided in this chapter captures some of the main ideas presented in the previous chapter, where the details and examples of the meaning of the framework were provided in greater depth. It is the evolving framework, grounded in the survey and interview data, which addresses the first research question. In considering the conceptions of EPSTs in the contexts of sampling, data and graphs, and probability situations, the framework provides structure for characterizing thinking about variation.

While the aspects and main dimensions within each aspect were hypothesized based on the work of other researchers with different subjects (e.g., Pfannkuch & Wild, 2001; Watson, 2000b), it remained an open question at the outset of this research whether or not EPSTs thought along the lines suggested by the initial framework posited at the end of Chapter Three. In conclusion, this research more deeply explored EPSTs' conceptions about variation, showed how those conceptions mapped into the evolving framework, and fleshed out richer detail in the framework itself.

Second Research Question

My second research question asked "How do EPST's conceptions of variation before an instructional intervention compare to those conceptions after the intervention?" This question was informed by the second part of Chapter Four, where the Emergent Framework proved useful for looking at individual conceptions of variation in considering the six cases' responses to a common subset of PostInterview

questions. I also used the framework to describe similarities among and differences between the six cases at the end of Chapter Four.

DS had some reasonable ideas about variation even at the start of the quarter, but in the PreInterview I was surprised that she considered Q3 ("Graph: 30") as showing real data. By the end of the quarter, along with the other cases, she knew that having such a narrow range as shown in Q3 was unrealistic. Whereas DS had talked earlier in the quarter about results not being "perfect", in the PostInterview she consistently talked more about expecting a range of results.

GP had not shown in the PreSurvey or the PreInterview that he had a very firm idea of what to expect or why, and he was prone to talking about physical causes of variation, especially in the way he might be able to draw out candies of a certain color. By the end of the quarter, GP had less emphasis on physical causes, and was giving more reasonable expectations, explanations, and interpretations. GP also repeatedly referred to experiences in class in his subsequent justifications. Finally, GP's manner in considering displays of variation started off with a heavy reliance on gesture, yet in the PostInterview he clearly had gained sophistication in his use of terminology to discuss graphs.

EM had an initial preoccupation with finding a mathematical formula. It seemed she thought that if she only could learn enough math, she could then make the correct predictions. By the end of the quarter, she expressed a more balanced view, considering proportional reasoning along with the variation she knew would be present. EM also made references to experiences done in class, and shifted in her

interpretation of variation by commenting on the PostInterview about influence of the number of trials on results (comments not made by EM in the PreInterview).

JM had a strong sense of proportional reasoning throughout the quarter, but was less tied to the idea of seeing average results in the PostInterview. He knew extreme results were possible, but developed in his sense of how unlikely those extreme values would be to occur. Also, while I think his appreciation for the physical causes of variation never went away, he mentioned these causes less frequently on the PostInterview. The biggest difference for JM, I believe, came from his own development of a sense of what really happens in situations where variation is inherent. Recall that JM had put all tens on Q9 ("Sixty Tosses" of the die) in the PreInterview. He never again made choices that exhibited such a lack of variation.

SP was emphatic in the PreSurvey and PreInterview that "Anything Can Happen" and "You Can Never Know" (other cases reflected these ideas, but none more so than SP). Consequently, SP had difficulty in making decisions about real or made-up data in the PreInterview, and she also had a marked lack of commentary about the expected value. She still did not explicitly mention average very much in the PostInterview, but her ranges for expectation were narrower. More importantly, she stopped talking about not knowing, or how anything could happen, and started giving more reasonable expectations and justifications.

Whereas RL was highly motivated by theoretical expectations at the outset of MET 2, over the quarter he increased in his appreciation of variation. For example, he countered his own inclination to offer only theoretical predictions for his expectations by offering ranges in the PostInterview. He also had a more sophisticated awareness

of influencing expectation and variation. In the PreInterview, he had conceded that even if individual results varied, the average of results should still match the expected value. However, he did move more in PostInterview towards the idea that means, medians, and modes also vary.

Using the CodeFrames bolstered my own analytic impression that by the end of the quarter, the six cases were closer together in terms of their reasoning than they were at the start of the quarter. Not only were they closer in agreement, they also each exhibited more mature reasoning. In particular, there were certain naïve features or responses for each case which had stood out in the early part of the quarter (on the PreSurvey or PreInterview) which were significantly diminished by the PostInterview, a change I attribute to the class interventions and also to the interview and survey tasks.

In summary, I saw some convergence when considering all six cases as they moved through the quarter. *Expectations* were more balanced: Predictions that were too narrow became wider, and wide ranges became narrower. Instead of "Anything Can Happen", extremes were seen as possible but unlikely. In *displaying* variation, graphs that were harder to decide as real or made-up became easier to adjudge. There was also better use of language in describing graphs, and it seemed that having experience with different graph types gave the cases more ways of evaluating and comparing graphs. The sense of *interpreting* variation also seemed more mature overall, with all cases having a reasonable view of how more trials influences expectation and variation.

One catalyst for changes in subject response was the class interventions, since the activities provided opportunities to explore and interact with sampling, data and graphs, and probability situations. Many subjects referred to something they had seen, heard, or thought of as a result of class interactions. Another catalyst for some changes in subject response was the survey or interview questions themselves. In other words, there was some self-learning that occurred as a direct result of interacting with the tasks I had provided in the surveys and interview scripts. The usefulness of research tasks addresses my third research question.

Third Research Question

My third research question asked "What tasks are useful for examining EPST's conceptions of variation in the contexts of sampling, data and graphs, and probability?" This question was informed by both the activities of the class interventions and the survey and interview tasks. I'll discuss some main highlights from the activities, surveys, and interviews, pointing out what made certain tasks useful.

Class Interventions

In the context of data and graphs, the "Four Questions" activity was useful for generating discussion about different measures of center and for bringing the importance of spread into the conversation. For instance, as we talked about what was "typical" for the number of pets in a household, some students wondered about the difficulty in saying what was typical without a consideration of both center and spread. The same tension between centers and spread arose when we discussed data for the "Body Measurements" activity. The latter activity also was useful for talking

about causes of variation, particularly for the repeated-measurements experiment in determining Matt's armspan. There were also opportunities to discuss the level of detail and subsequent usefulness of different types of graphs when the class engaged in the "Four Questions" and "Body Measurements" activities.

In the contexts of sampling and probability, there were three ways in which all the activities ("Known Mixture", "Unknown Mixture", "Cereal Boxes", and the "River Crossing Game") were extremely useful for examining conceptions of variation. The first way is that the activities all allowed for initial discussion to bring out *what* subjects expected and *why*. Second, the activities all gave students hands-on opportunities to see for themselves what kind of variation really results from sampling and probability situations. Third, the activities all lent themselves to using computer simulations to show what happens with extremely large numbers of samples.

For example, both the "Known Mixture" and the "River Crossing Game" involved students making predictions ahead of time for what they thought might happen, and many students had some initial idea of what the underlying distribution looked like. As we talked in class about what we might expect, some students suggested that the mode would be at the expected value, while others mentioned the kind of range they thought might result. I think it is very important for teacher educators to thoroughly discuss predictions ahead of time and not just jump into activities to see the actual results. A huge opportunity will be lost if students are not asked ahead of time what they expect, and why, in situations involving variability. The pedagogical payoff comes from relating the post-activity discussions back to the

ideas prompted in the pre-activity discussion. Especially when records of initial predictions and actual results are posted up in the class for everyone to see, it becomes easy for students to reflect on differences and similarities they notice.

The main reason I think the physical data gathering is so important for students is because there seems to be some cognitive need to see for themselves what will happen by actually doing the experiment with their own hands. Piaget (1975) wrote about how children at the stage of concrete operational thought develop mathematical ideas as they engage and reflect upon activity in a tangible environment. Piaget's ideas about the importance of concrete operations transfers to the adult learner, and thus it is important to offer physical experiences in data-gathering not only to children, but to the prospective teachers of children. The experiences were useful for convincing EPSTs, for instance, how they really don't usually get the same result each time, and that even if they try to roll the dice the same way they'll still see variation in their results. Making the graphs of results by hand encouraged the students pay attention to different elements of the distribution. The physical data collection also paved the way for understanding what the computer simulations were accomplishing.

The usefulness of the computer simulations were apparent from the way that so many subjects commented on them afterwards. They noted how many trials it took to get extreme results, and we also called attention to the changes in shape of the distribution of cumulative results as we did more and more samples. I think it is critical to do and then discuss the hands-on data gathering before doing any computer simulations, because otherwise some students may not fully appreciate what is going on with the computer displays. The MET 2 class slowly aggregated samples by first

doing experiments in small groups, then combining results from a few groups, and then looking at classwide data. Whether we were looking at tens or hundreds of samples, it was only after we had thoroughly discussed results gained from our hands-on experiments that we turned to the computer.

Survey Tasks

Most of the survey tasks were either direct copies of or very similar to tasks used by other researchers, because those tasks had already proven useful in examining conceptions of variation with precollege students. For example, the "One Sample", "Several Samples", and "Six Samples" questions for drawing candies from a jar had all been used in prior research, and I also applied those questions to flips of a fair coin and spinner scenarios. There were a few tasks that I created or adapted for survey use. I'll highlight some of those tasks from the Data & Graphs PostSurvey, discussing what made them useful for examining conceptions of variation held by EPSTs.

The rainfall tasks on the Data & Graphs PostSurvey were useful for drawing out students' ideas about causes of variation. Having data presented in two types of graphs was useful for having students attend to different elements of the distribution. The rainfall data was presented in both boxplots and bar charts, and I could see how the height of bars made a visual impression on some students while the width of the box was a focus for some other students.

I also phrased a couple of the rainfall questions so that students reacted to an argument given by some hypothetical person, such as, "Zain said Columbus was rainier because the average monthly rainfall was higher than Portland." I found that this "React to an Argument" style of question provided a good springboard for

students to tell me what they thought. In fact, when I asked the React-to-an-Argument type of questions I tended to get more information than when I asked an open-ended question. For example, when I asked an open-ended question about traffic deaths ("How do traffic deaths rates in the South compare with those in the Northeast?") often I just got responses telling me how the South had a higher average. When I asked an open-ended question about which city the students thought was rainier and why, I got some very good answers involving different elements of the distributions but I also got many more responses that just expressed personal opinions about how students felt about rain. In contrast, when I asked students to react to someone else's argument based on extreme values (such as how Portland had the highest rainfall), I read lengthier responses that showed greater detail about what the students were thinking about the theme of range of extremes. The "React to an Argument" style of questioning has also been used by other researchers to gather data about how students think about probability and statistics (e.g., Jacobs, 1997; Watson et. al., 2002).

The question in the Data & Graphs PostSurvey about generating a graph to show daily rainfall based on knowing the monthly average was motivated by the approach of Mokros & Russel (1995), and was extremely useful for two reasons. First, the responses gave me some idea of the students' graph sense, because although I provided two labeled axes and scaled the horizontal axis, I did not specify what kind of graph they should use and I did not scale the vertical axis. I was surprised at the variety of graph types the students used, and that some of their graphs were better-suited to showing daily variation then others. Second, I was able to see the kind of variation they expected in this situation, and in some cases there were big surprises.

For example, some graphs showed rain every single day, in varying amounts. Some graphs showed no rain for most days, and a couple of graphs showed exactly the same amount of rain for each day. In retrospect, it would have been a great idea to take a document camera and show the class some of their classmates' graphs and ask what they thought about the amount of variation shown.

Interviews

The interviews, like the surveys, also contained some tasks that had already proven useful in other research (e.g., Watson et. al., 2002; Shaughnessy, Ciancetta, & Canada, 2003). Tasks like "One Sample", "Several Samples", and "Six Samples" let students make their own predictions and justify their choices. Tasks like "Compare Lists" let students react to given predictions. In an interview setting, I found that there are two key features that make a task particularly useful. One feature is how easily the subject is able to get engaged in the task. If a subject readily understands the nature of the task, finds it interesting, and is able to talk about it with little prompting, then it is easier to gather data from that subject. Another feature is the quality of the data gathered. That is, if the subject is offering thoughts germane to the research, then the data is useful. Thus, copious and relevant input from the subjects were two hallmarks of useful tasks, and I'll profile just a few of the more interesting interview tasks that I had either created or substantially modified based on questions used in other research.

The "MAX Wait-Times" question in the PreInterview was useful for highlighting the tension between centers and spread. Since the data sets had identical means and medians some subjects were initially attracted to the claim that there was no real difference in wait-times. Other subjects focused right away on the different

spreads of the two data sets, and talked about how the average doesn't give an accurate picture of the data. I had included a React-to-an-Argument type of probe in the "MAX Wait-Times" situation, but left the line of questioning more open-ended in the similar "Muffin Weights" question on the PostInterview. Also, the two "Muffin Weights" data sets did not have equal averages, and the bakery with more spread had a higher mean and mode. Subjects seemed very willing to discuss the graphs in both "MAX Wait-Times" and "Muffin Weights", and in both questions they volunteered some detailed information about what they were thinking in terms of the distribution. In "Muffin Weights", I was able to see how subjects interacted with boxplots versus dotplots, since I used both types of graphs in that task. The "MAX Wait-Times" question was later modified into a "Movie Wait-Time" question which was then used in research with middle and high school students.

Several tasks on both interviews had a common Real-versus-Fake dynamic, including the "Graph: 30", "Graph: 300", and "Likelier Graph?" questions. I varied the specific wording on the different questions, but the basic idea was always the same: Did subjects think a graph reflected real or made-up data? Every time I asked any subject a question having a Real-versus-Fake dynamic, the subject seemed to have no trouble talking about what he or she was thinking. That is not to say that all subjects were quick to decide, because several subjects wrestled at length even in coming to a decision of no confidence. I thought it was important to include "no confidence" when asking students what they thought was most likely, because otherwise they may have felt compelled to make a choice between the two other choices of "real" or "made-up".

I combined a Real-versus-Fake dynamic with several React-to-an-Argument probes in PostInterview Q12, which I nicknamed "Compare Comments." The probes specifically asked students to react to comments about different elements of the distribution of the graph under question. My subjects found it very easy to be assertive when reacting to given arguments, and their responses typically addressed themes of the evolving framework. For example, the expected value of the problem was 25 blacks, and one given argument was how "Keith argued that something was wrong with the experiment because no one got exactly 25 out of 50 landing on black." Here was RL's reaction:

> RL: Well, I don't think that –just because somebody, nobody got 25, that seems to me a little bit nit-picky, uh, because you're not – That's adherence , that's too close adherence to this principle of "It's theoretical, and therefore that's what I expect to see" And what Keith is not appreciating, in fact, I think a couple people here are overlooking the fact that they spun it 20 [sets of 50 spins each]… But ONLY 20 sets. And so, do it 10,000, see, you know? Come back and talk about that.

I noticed a corrective tone in RL's response as he was telling me what Keith was "not appreciating", and RL offered some valid counter-arguments of his own. I think that "React-to-an-Argument" questions, while directed more by the researcher and therefore less open-ended initially, definitely generate useful data and seem to make it very easy for subjects to say what they think.

The Real-versus-Fake questions described so far in this section all involved graphs, and were inspired by questions used in previous research (e.g., Watson et. al., 2002; Shaughnessy et. al., 2004). However, the first two die-tossing questions in the PreInterview did not involve graphs but still had a Real-versus-Fake dynamic.

PreInterview Q9 ("One Sample") and Q10 ("Who Cheated?") were powerful questions because of the cognitive conflict they helped invoke. The key to the two questions' strength, I believe, lay in the Before-and-After sequencing of the questions. That is, in Q9 the subjects were asked to imagine what results might actually occur before an experiment was to take place. In Q10, subjects were presented with results that were reported after the experiment had supposedly been done. Regardless of whether a subject had put all tens or not in Q9, every single subject seemed to evaluate the entire situation differently in Q10. It was as if the question itself took on a new level of importance once we got to Q10 and I suggested to my subjects that they would have to decide if their hypothetical students were cheating or not.

Limitations of Research

There are two limitations regarding this research that I want to mention. One concerns the themes within the framework, and the other concerns the class environment.

The themes of the framework are useful for looking at EPSTs' conceptions of variation, but are not guaranteed to easily characterize all possible responses. One example of a type of response that did not easily fit into the framework concerned levels of surprise. On the PreInterview, I asked a series of questions based on Truran's (1994) research tasks, asking subjects about a series of outcomes to find out what was surprising. At first I had considered adding "Concerning Levels of Surprise" to go along with the other themes listed in [1A] for *what* was *expected*. However, in the PostInterview, a case used the language of surprise in a way that suggested a reason *why* expectations were held, and it seemed that "surprising" was linked to possibilities

and likelihood. Thus, it was unclear whether responses involving a sense of surprise fit more naturally with *what* was *expected* or with *why*. Truran's idea of a series of questions leading to a sort of "surprise threshold" helps reveal what is or is not expected, but at the same time the notion of surprise also can offer a form of justification. The dilemma is much akin to expecting results to vary because there should be variation: The way the students phrase their response and the context of the question give clues about what theme best fits their idea. Thus, some of the themes within the framework could use some additional sharpening in definition.

There also may be additional conceptions not addressed by the framework. As a first look at EPSTs conceptions of variation, the framework has much to offer, but I suggest further possible refinements in the next section.

Another possible limitation of the research concerns the class environment. The culture of the MET 1 and MET 2 classes were largely defined by the in-class activities, group interactions, and spirit of student-driven inquiry. Almost all the students who participated in the research had taken MET 1 at the same university where the research was conducted. Over half of the students completing the surveys – and all of the case studies – had taken the prerequisite course with the same instructor, Steve, whose teaching exemplified the class culture earlier described. Thus, my sense of the students was that they were experienced in describing their own reasoning, communicating how they were thinking both verbally and in writing. However, it is not clear what replication of results would be found among other EPSTs at other universities, especially given the considerable variation among teacher preparation

programs. I would expect the conceptions of other EPSTs to fall within the framework, but further study is warranted.

Implications for Research and Teaching

There are three areas for which I recommend future research relating to the continued improvement of preservice teacher education about variation. One area concerns the refinement and testing of the framework; a second area concerns comparing preservice teachers' conceptions with the conceptions of school students; a third area concerns the curriculum for teacher preparation.

Refinement

To further sharpen some of the definitions of the themes within the framework, research tasks should be crafted to tease apart finer shades of meaning. For example, in comparing data sets, sometimes students referred to variation as a synonym for range, and sometimes variation meant the distribution of data within the range. It became problematic when the students had alternate meanings within the same response, and some new tasks or new lines of questioning could be designed to clarify these problematic situations. Also, using some of the survey items on a large scale with preservice teachers across several universities would accomplish two useful purposes. First, the overall utility of the framework could be tested on a stronger quantitative basis than was offered in this research, and one could begin to investigate the generalization of the application of the framework. Second, interactions within the framework could be examined with greater clarity. For instance, are students with stronger *interpretations* likelier to have better *expectations*? Do students who make reasonable *comparisons* of graphs also produce reasonable graphs themselves? There

are many questions suitable to a more quantitative study, given that this research has provided a critical first step towards identifying the important aspects of variation and what comprises those aspects.

Comparisons Across Age Levels

Previous research has looked at or is looking at conceptions of variation held by elementary, middle, and high school students. My research looked at prospective teachers of students. I recommend studies designed to compare the conceptions of students and their prospective teachers. A possible benefit of such a comparison could be the design of better curricula for classroom teaching, since such curricula would be informed not only by a sense of student conceptions, but also by preservice teachers' conceptions.

Curriculum Development

A study designed specifically as a teaching experiment would be appropriate. This research has pointed out relevant aspects to focus upon. This research has also laid out some useful interventions to consider. However, to actually measure effectiveness in a classroom setting it would require additional research that aims more at the teaching and learning within a class. Steve is a seasoned MET 2 teacher, and Matt and I were experienced in working with class interventions for variation at the middle and high school level. Since all three of us had a hand in the MET 2 interventions, it is safe to say that the subjects in this research had a fairly unique experience. Regarding the teaching and learning about variation, how do the actions and background of the college instructor shape the dialogue and experiences of the

preservice teacher? Research designed along the lines of a teaching experiment could address that question, and others such as: What are the most effective ways to construct a class intervention about variation? How much computer simulation is appropriate, and how should those simulations be designed? There is much more that research can contribute to finding optimal ways to structure courses for preservice teachers, especially concerning probability and statistics.

This research already provides a number of suggestions for teachers of teachers of mathematics. The research implies that it is not sufficient to merely address normative measures such as range and standard deviation in order to address conceptions of variation. Preservice teachers need to have opportunities to address all three aspects: *expecting*, *displaying,* and *interpreting* variation. They need these opportunities within different contexts, such as sampling, data and graphs, and probability. With students like SP or GP, for example, it would have been easy to assume they had an overall weak appreciation for variation at the outset of the course, based on some unreasonable expectations or justifications which they provided on the PreSurvey and in the PreInterview. However, because the instruments varied in context, I was able to see, in the case of GP for example, that while he had some questionable ideas about *sampling*, he had a natural inclination towards considering variation in the context of *data and graphs*. Also, while his language in discussing graphs in the PreInterview was less sophisticated, he made heavy use of gesture to convey some very reasonable ideas. By attending to different contexts and ways of expressing ideas, a better picture emerges of what preservice teachers can and do understand about variation.

Concluding Comments

Ultimately, it is precisely what EPSTs *do* understand about variation that sets this research apart. Finding out what learners *don't* know about probability and statistics is one approach to research, exemplified by earlier studies about intuition and misconceptions, but the focus for this research has been on what learners *do* know. My research adds to the literature in the area of statistical education by offering an in-depth exploration of the conceptions of EPSTs about variation, along with a detailed framework for characterizing their conceptions. Finding out the conceptions of variation held by EPSTs lays the groundwork for improved instruction at the college level, in turn resulting in better experiences for children at the schools where the EPSTs eventually serve.

REFERENCES

Australian Education Council. (1991). A National Statement on Mathematics for Australian Schools. Carlton, Victoria: Curriculum Corporation.

Ball, D. (1990). The mathematical understandings that prospective teachers bring to teacher education. The Elementary School Journal, 90, 449-465.

Ball, D. (1993). Halves, pieces, and twoths: Constructing and using representational contexts in teaching fractions. In T. Carpenter, E. Fennema, & T. Romberg (Eds.), Rational Numbers: An Integration of Research , (pp.157-195). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Ball, D., & McDiarmid, G. (1988a). Research on teacher learning: Studying how teachers' knowledge changes. Action in Teacher Education, 10, 17-23.

Ball, D., & McDiarmid, G. (1988b). The subject-matter preparation of teachers. In W.R. Houston (Ed.), Handbook of Research on Teacher Education (pp. 437-449). New York: Macmillan.

Bar-Hillel, M. (1982). Studies of representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgment Under Uncertainty: Heuristics and Biases. New York, NY: Cambridge University Press.

Batanero, C., Godino, J., Valecillos, A., Green, D., & Holmes, P. (1994). Errors and difficulties in understanding elementary statistical concepts. International Journal of Mathematical Education in Science and Technology, 25 (4), 527-547.

Batanero, C., Green, D., & Serrano, L. (1998). Randomness, its meanings and educational implications. International Journal of Mathematical Education in Science and Technology, 29 , 113-123.

Batanero, C., & Serrano, L. (1999). The meaning of randomness for secondary school students. Journal for Research in Mathematics Education, 30 , 558-567.

Becker, H. (1991). Theory: The necessary evil. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Begg, A. (1995). Making sense when learning mathematics. In J. Neyland (Ed.), Mathematics Education: A Handbook for Teachers, Vol. 2 , (pp. 70-76). Wellington, NZ: Wellington College of Education.

Beltrami, E. (1999). What is Random? Chance and Order in Mathematics and Life. New York, NY: Springer - Verlag.

Bennett, D. (1998). Randomness. Cambridge, MA: Harvard University Press.

Bennett, A., & Foreman, L. (1991). Visual Mathematics Course Guide, Volume II. Salem, OR: Math Learning Center.

Bennett, A., & Nelson, L.T. (2001). Mathematics for Elementary Teachers: A Conceptual Approach, 5th Edition. Boston, MA: McGraw-Hill.

Best, J. & Kahn, J. (1998). Research in Education (8th ed.). Needham Heights, MA: Allyn & Bacon.

Bidwell, J., & Clason, R. (1970). Readings in the History of Mathematics Education. Washington, DC: National Council of Teachers of Mathematics.

Borko, H., Eisenhart, M., Brown, C., Underhill, R., Jones, D., & Agard, P. (1992). Learning to teach hard mathematics: Do novice teachers and their instructors give up too easily? Journal for Research in Mathematics Education, 23, 194-222.

Bright, G., & Friel, S. (1998). Graphical representations: Helping students interpret data. In S. Lajoie (Ed.), Reflections on Statistics: Learning, Teaching, and Assessment in Grades K-12, (pp. 63-88). Mahwah, NJ: Lawrence Erlbaum Associates.

Callingham, R. (1997). Teachers' multimodal functioning in relation to the concept of average. Mathematics Education Research Journal, 9 (2), 205-224.

Callingham, R., Watson, J., Collis, K., & Moritz, J. (1995). Teacher attitudes towards chance and data. In B. Atweh & S. Flavel (Eds.), Proceedings of the Eighteenth Annual Conference of the Mathematics Education Research Group of Australasia , (pp. 143-150). Darwin, NT: Mathematics Education Research Group of Australasia.

Cambridge Conference on School Mathematics. (1963). Goals for School Mathematics: The Report of the Cambridge Conference on School Mathematics. Boston, MA:Houghton-Mifflin.

Cobb, P. (1994). Where is the mind? Constructivist and sociocultural perspectives on mathematical development. Educational Researcher, 23 (7) , 25-27.

Cobb, P. (2000). Conducting teaching experiments in collaboration with teachers. In R. Lesh & A. Kelly (Eds.), Handbook of Research Design in Mathematics and Science Education (pp. 307-333). Mahwah, NJ: Lawrence Erlbaum Associates.

Cobb, G., & Moore, D. (1997). Mathematics, Statistics, and Teaching. American Mathematics Monthly, 104 (9), 801-824.

Cobb P., & Yackel, E. (1996). Constructivist, emergent, and sociocultural perspectives in the context of developmental research. Educational Psychologist, 31 (3/4) , 175-190.

Cobb, P., Yackel, E., & Wood, T. (1992). A constructivist alternative to the representational view of mind in mathematics education. Journal for Research in Mathematics Education, 23 (1), 2-33.

Cooney, T. (1994). Research and teacher education: In search of a common ground. Journal for Research in Mathematics Education, 25 (6), 608-636.

Creswell, J. (1998). Qualitative inquiry and research design: Choosing among five traditions. Thousand Oaks, CA: Sage Publications.

Curcio, F. (1987). Comprehension of mathematical relationships expressed through graphs. Journal for Research in Mathematics Education, 18 (5), 382-393.

Darling-Hammond, L. (1998). Teachers and teaching: Testing policy hypothesis from a National Commission report. Educational Researcher, 27, 5 - 15.

Davis, P., & Hersh, R. (1986). Descarte's Dream: The World According to Mathematics. San Diego, CA: Harcourt Brace Jovanovich.

Department for Education (England and Wales). (1995). Mathematics in the National Curriculum. London, UK: Author.

Eisenhart, M. (1991). Conceptual frameworks for research circa 1991: Ideas from a cultural anthropologist; Implications for mathematics education researchers. In R. Underhill (Ed.), Proceedings of the 13th Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, Vol. 1 , (pp. 202-219). Blacksburg, VA: Virginia Polytechnic Institute and State University.

Eisenhart, M., Borko, H., Underhill, R., Brown, C., Jones, D., & Agard, P. (1993). Conceptual knowledge falls through the cracks: Complexities of learning to teach mathematics for understanding. Journal for Research in Mathematics Education, 24, 8-40.

Erlwanger, S. (1973). Benny's conception of rules and answers in IPI mathematics. Journal of Children's Mathematical Behavior, 1 (2), 7-25.

Ernest, P. (1991). The Philosophy of Mathematics Education. Hampshire, U.K.: Falmer Press.

Ernest, P. (1996). Varieties of constructivism: A framework for comparison. In L. Steffe, P. Nesher, P. Cobb, G. Goldin, & B. Greer (Eds.), Theories of Mathematical Learning (pp. 335-350). Mahwah, NJ: Lawrence Erlbaum Associates.

Ernest, P. (1997). The epistemological basis of qualitative research in mathematics education: A postmodern perspective. In A. Teppo (Ed.), Qualitative Research Methods in Mathematics Education (pp. 22-39). Reston, VA: National Council of Teachers of Mathematics.

Ernest, P. (1998). Social Constructivism as a Philosophy of Mathematics. Albany, NY: State University of New York Press.

Even, R. (1993). Subject-matter knowledge and pedagogical content knowledge: Prospective secondary teachers and the function concept. Journal for Research in Mathematics Education, 24, 94 - 116.

Falk, R. (1983). Experimental models for resolving probabilistic ambiguities. In R. Hershkowitz (Ed.), Proceedings of the Seventh International Conference for the Psychology of Mathematics Education, (pp. 319-325). Rehovot, Israel: Weizmann Institute of Science.

Fennema, E., & Franke, M. (1992). Teachers' knowledge and its impact. In D. Grouws (Ed.), Handbook of Research on Mathematics Teaching an Learning, (pp. 147-163). New York: Macmillan.

Fischbein, E., & Gazit, A. (1984). Does the teaching of probability improve probabilistic intuitions? Educational Studies in Mathematics, 15 , 1-24.

Fischbein, E., Pampu, I., & Minzat, I. (1975). The child's intuition of probability. In E. Fischbein (Ed.), The Intuitive Sources of Probabilistic Thinking in Children (pp. 156-174). Dordrecht: Reidel.

Fischbein, E., & Schnarch, D. (1997). The evolution with age of probabilistic, intuitively based misconceptions. Journal for Research in Mathematics Education, 28 (1), 96-106.

Flevares, L, & Perry, M. (2001). How many do you see? The use of nonspoken representations in first-grade mathematics lessons. Journal of Educational Psychology, 93 (2), 330-345.

Fontana, A., & Frey, J. (1994). Interviewing: The art of science. In N. Denzin & Y. Lincoln (Eds.), Handbook of Qualitative Research (pp. 361-376). Thousand Oaks, CA: Sage Publications.

Foreman, L., & Bennett, A. (1995). Visual Mathematics, Course I. Salem, OR: The Math Learning Center.

Ford, J. (1983, April). How random is a coin toss? Physics Today, 40-47.

Fosnot, C. (1996). Constructivism: A psychological theory of learning. . In C. Fosnot (Ed.), Constructivism: Theory, Perspectives, and Practice (pp. 8-33). New York, NY: Teachers College.

Fox, R. (1997). Constructivist views of learning. In P. Preece (Series Ed.) & R. Fox (Issue Ed.), Perspectives on Constructivism: Perspectives 56 (pp. 3-16). Exeter, UK: University of Exeter School of Education.

Fraenkel, J. & Wallen, N. (2000). How to Design and Evaluate Research in Education (4th ed.). Boston, MA: McGraw Hill.

Franke, M. (2000). How much can we accomplish? Elementary mathematics methods revisited. Mathematics Education Dialogues, 4 (1), 9.

Friel, S., & Bright, G. (1996). Building a theory of graphicacy: How do students read graphs? Paper presented at the American Educational Research Association, NewYork, NY.

Friel, S., & Bright, G. (1998). Teach-Stat: A model for professional development in data analysis and statistics for teachers K-6. In S. Lajoie (Ed.), Reflections on Statistics: Learning, Teaching, and Assessment in Grades K-12, (pp. 89-117). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Friel, S., Bright, G., Frierson, D., & Kader, G. (1997). A framework for assessing knowledge and learning in statistics. In I. Gal & J. Garfield (Eds.), The Assessment Challenge in Statistics Education , Amsterdam, The Netherlands: IOS Press.

Gal, I., & Garfield, J. (1997). Curriculum goals and assessment challenges in statistics education. In I. Gal & J. Garfield (Eds.), The Assessment Challenge in Statistics Education (pp. 1- 13). Amsterdam, The Netherlands: IOS Press.

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. Journal for Research in Mathematics Education, 19 (1), 44-63.

Gawronski, J. & McLeod, D. (1980). Probability and Statistics: Today's Ciphering? In M.M. Lindquist (Ed.), Selected Issues in Mathematics Education, (pp. 82-89). United States: McCutchan Publishing.

Goldin, G. (2000). A scientific perspective on structured, task-based interviews in mathematics education research. . In R. Lesh & A. Kelly (Eds.), Handbook of Research Design in Mathematics and Science Education (pp. 517-545). Mahwah, NJ: Lawrence Erlbaum Associates.

Goldin-Meadow, S., Kim, S., & Singer, M. (1999). What the teacher's hands tell the student's mind about math. Journal of Educational Psychology, 91 (4), 720-735.

Green, D. (1983). A Survey of Probability Concepts in 3000 Pupils Aged 11-16 Years. In V. Barnett, G. Constable, D. Grey, & P Holmes (Eds.) Proceedings of the First International Conference on Teaching Statistics, (pp. 766-783). Sheffield, U.K.: Teaching Statistics Trust.

Grossman, P., Wilson, S., & Shulman, L. (1989). Teachers of substance: Subject matter knowledge for teaching. In M. Reynolds (Ed.), The Knowledge Base for the Beginning Teacher, 23-36. New York: Pergamon Press.

Harvard, G. (1997). The key ideas of Vygotsky and their implications for teaching and schooling. In P. Preece (Series Ed.) & R. Fox (Issue Ed.), Perspectives on Constructivism: Perspectives 56 (pp. 23-37). Exeter, UK: University of Exeter School of Education.

Herscovics, N. (1996). The construction of conceptual schemes in mathematics. In L. Steffe, P. Nesher, P. Cobb, G. Goldin, & B. Greer (Eds.), Theories of Mathematical Learning (pp. 351-379). Mahwah, NJ: Lawrence Erlbaum Associates.

Horvath, J., & Lehrer, R. (1998). A model-based perspective on the development of children's understanding of chance and uncertainty. In S. Lajoie (Ed.), Reflections on Statistics: Learning, Teaching, and Assessment in Grades K-12 , (pp. 121-148). Mahwah, NJ: Lawrence Erlbaum Associates.

Huberman, A., & Miles, M. (1994). Data management and analysis methods. In N. Denzin & Y. Lincoln (Eds.), Handbook of Qualitative Research (pp. 428-444). Thousand Oaks, CA: Sage Publications.

Jacobs, V. (1997). Children's understanding of sampling in surveys. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Janvier, C. (1996). Constructivism and its consequences for training teachers. In L. Steffe, P. Nesher, P. Cobb, G. Goldin, & B. Greer (Eds.), Theories of Mathematical Learning (pp. 449-463). Mahwah, NJ: Lawrence Erlbaum Associates.

Jones, G., Thornton, C., Langrall, C., Mooney, E., Perry, B., & Putt, I. (2000). A framework for characterizing children's statistical thinking. Mathematical Thinking and Learning, 2 (4), 269-307.

Kac, M. (1983). What is random? American Scientist, 71, 405-406.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. Cognitive Psychology, 3 , 430-451.

Kolata, G. (1986). What does it mean to be random? Science, 231, 1068 -1070.

Konold, C. (1989). Informal conceptions of probability. Cognition and Instruction, 6 (1), 59-98.

Konold, C. (1995). Confessions of a coin flipper and would-be instructor. American Statistician, 49 (2), 203-210.

Konold, C., Pollatsek, A., Well, A., Lohmeier, J., Lipson, A. (1993). Inconsistencies in students' reasoning about probability. Journal for Research in Mathematics Education, 24 (5), 392-414.

Kozulin, A. (1986). Vygotsky in context. Preface to Vygotsky, L., Thought and Language. Cambridge, MA: The MIT Press.

Kuzmak, S., & Gelman, R. (1986). Young children's understanding of random phenomena. Child Development, 57, 559-566.

Lajoie, S., & Romberg, T. (1998). Identifying an agenda for statistics instruction and assessment in K-12. In S. Lajoie (Ed.), Reflections on Statistics: Learning, Teaching, and Assessment in Grades K-12, (pp. xi-xxi). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Landwehr, J. & Watkins, A. (1995). Exploring Data (Revised ed.). United States of America: Dale Seymour Publications.

Lappan, G., Fey, J., Fitzgerald, W., Friel, S., & Phillips, E. (1998). Connected Mathematics (Grade 9): Samples and Populations. White Plains, NY: Dale Seymour Publications.

Lehrer, R., & Franke, M. (1992). Applying personal construct psychology to the study of teachers' Knowledge of fractions. Journal for Research in Mathematics Education, 23, 233-241.

Leinhardt, G., & Smith, D. (1985). Expertise in mathematics instruction: Subject matter knowledge. Journal of Educational Psychology, 77, 247-271.

Leon, M., & Zawojewski, J. (1990). Use of the Arithmetic Mean: An Investigation of Four Properties (Issues and Preliminary Results). Paper presented at the Third International Conference on Teaching Statistics (ICOTS III), Dunedin, New Zealand.

Lester, F. (1991). The nature and purpose of research in mathematics education: Ideas prompted by Eisenhart's plenary address. In R. Underhill (Ed.), Proceedings of the 13th Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, Vol. 1 , (pp. 193-201). Blacksburg, VA: Virginia Polytechnic Institute and State University.

Lockwood, A. (1998). Megan Loef Franke: Professional development & mathematical understanding. Principled Practice in Mathematics & Science Education, 2, 3 - 6.

Loosen, F., Lioen, M., & Lacante, M. (1985). The standard deviation: Some drawbacks of an intuitive approach. Teaching Statistics, 7 (1), 29-39.

Marshall, C. & Rossman, G. (1995). Designing Qualitative Research (2nd ed.). Thousand Oaks, CA: Sage Publications.

May, M. (1997). What is random? American Scientist, 85 (3), 222 -223.

McDiarmid, G., Ball, D., & Anderson, C. (1989). Why staying one chapter ahead doesn't really work: Subject-specific pedagogy. In M. Reynolds (Ed.), The Knowledge Base for the Beginning Teacher, (pp. 193-205). New York: Pergamon Press.

McMillan, J. & Schumacher, S. (1997). Research in Education (4th ed.). New York, NY: Longman.

Mellissinos-Lernhardt, M., (1999). What the Mean Does (and Does Not) Tell Students About a Distribution. Paper presented at the Research Presessions of the 77th Annual Meeting of the National Council of Teachers of Mathematics.

Mellissinos, M., Ford, J., & McLeod, D. (1997). Student understanding of statistics: Developing the concept of distribution. In J. Dossey & J. Swafford (Eds.), Proceedings of the 19th Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education , Bloomington: IL.

Merriam, S. (1998). Qualitative Research and Case Study Applications in Education. San Francisco, CA: Jossey-Bass.

Metz, K. (1998). Emergent ideas of chance and probability in primary-grade children. In S. Lajoie (Ed.), Reflections on Statistics: Learning, Teaching, and Assessment in Grades K-12 , (pp. 149-173). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Metz, K. (1999). Why sampling works or why it can't: Ideas of young children engaged in research of their own design. In F. Hill & M. Santos (Eds.), Proceedings of the 21st Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, Vol. 2 , (pp. 492-498). Columbus, OH.

Miles, M., & Huberman, A. (1994). Qualitative Data Analysis: An Expanded Sourcebook (2nd ed.). Thousand Oaks, CA: Sage Publications.

Ministry of Education. (1992). Mathematics in the New Zealand Curriculum. Wellington, NZ: Author.

Mokros, J., & Russell, S. (1995). Children's concepts of average and representativeness. Journal for Research in Mathematics Education, 26 (1), 20-39.

Mokros, J., Russell, S., Weinberg, A., & Goldsmith, L. (1990). What's Typical? Children's Ideas about Average. Paper presented at the Third International Conference on Teaching Statistics (ICOTS III), Dundedin, New Zealand.

Moore, D. (1990). Uncertainty. In L. Steen (Ed.), On the Shoulders of Giants (pp. 95-137). Washington, DC: Academy Press.

Moore, D. (1991). Statistics: Concepts and Controversies (3rd ed.). New York, NY: W.H. Freeman and Company.

Moore, D. (1997). New pedagogy and new content: The case of statistics. International Statistical Review, 65 (2), 123-165.

Moritz, J., & Watson, J. (1997). Graphs: Communication lines to students? In F. Bidduch & K. Carr (Eds.), Proceedings of the 20th Annual Conference of the Mathematics Education Research Group of Australasia , (pp. 344- 351). Rotorua, NZ: MERGA.

National Council of Teachers of Mathematics. (1970). A History of Mathematics Education in the United States and Canada. Washington, DC: NCTM.

National Council of Teachers of Mathematics. (1980). An Agenda for Action: Recommendations for School Mathematics of the 1980s. Reston, VA: Author.

National Council of Teachers of Mathematics. (1989). Curriculum and Evaluation Standards for School Mathematics. Reston, VA: NCTM.

National Council of Teachers of Mathematics. (1991). Professional standards for teaching mathematics. Reston, VA: NCTM.

National Council of Teachers of Mathematics. (2000). Principles and Standards for School Mathematics. Reston, VA: NCTM.

Neyland, J. (1995). Beliefs and values in mathematics education: An outline of Ernest's model. In J. Neyland (Ed.), Mathematics Education: A Handbook for Teachers,Vol. 2 , (pp. 139-149). Wellington, NZ: Wellington College of Education.

Noddings, N. (1990). Constructivism in mathematics education. In R. Davis, C. Maher, & N. Noddings (Eds.), Constructivist Views on the Teaching and Learning of Mathematics (pp. 7-18). Reston, VA: National Council of Teachers of Mathematics.

Olive, J. (1991). Logo programming and geometric understanding: An in-depth study. Journal for Research in Mathematics Education, 22 (2), 90-111.

Pagels, H. (1982). The Cosmic Code: Quantum Physics as the Language of Nature. New York, NY: Simon & Schuster.

Patton, M. (2001). Qualitative evaluation and research methods. (3rd ed.). Thousand Oaks, CA: Sage Publications.

Pfannkuch, M. (1997). Statistical thinking: One statistician's perspective In F. Bidduch & K. Carr (Eds.), Proceedings of the 20th Annual Conference of the Mathematics Education Research Group of Australasia , (pp. 406-413). Rotorua, NZ: MERGA.

Pfannkuch, M., & Wild, C. (1998). Statistical thinking and statistical practice: Themes gleaned from professional statisticians. Unpublished manuscript.

Pfannkuch, M., & Wild, C. (2001). What do we know about statistical thinking? Overview of statistical thinking, a literature review. In C. Reading (Ed.), Background Readings for the Second International Research Forum on Statistical Reasoning, Thinking, and Literacy , Armidale, Australia.

Piaget, J. (1937). The Language and Thought of a Child. (M. Gabain, Trans.). New York, NY: Harcourt, Brace, and Company.

Piaget, J. (1964). Development and learning. Journal of Research in Science and Teaching, 2, 176-186.

Piaget, J. (1975). Piaget's theory (G. Cellerier & J. Langer, trans.). In P.B. Neubauer (Ed.), The process of child development, (pp. 164-212). New York: Jason Aronson.

Piaget, J., & Inhelder, B. (1975). The Origin of the Idea of Chance in Children. (L. Leake, P. Burrell, & H. Fishbein, Trans.). Toronto: W. W. Norton & Company, Ltd. (Original work published in 1951).

Pollatsek, A., Lima, S., & Well, A. (1981). Concept of computation: Students' understanding of the mean. Educational Studies in Mathematics, 12, 191-204.

Porter, A., & Brophy, J. (1988). Synthesis of research on good teaching: Insights from the Work of the Institute for Research on Teaching. Educational Leadership, 45 (8), 74-85.

Quinn, R. (1997). Effects of mathematics methods courses on the mathematical attitudes and content knowledge of preservice teachers. The Journal of Educational Research, 91, 108-113.

Reading, C., & Pegg, J. (1996). Exploring understanding of data reduction. In L. Puig & A. Gutierrez (Eds.), Proceedings of the 20th Conference of the International Group for the Psychology of Mathematics Education, Vol. 4 , (pp. 187-194). Valencia: Spain.

Reading, C., & Shaughnessy, J. (2000). Students' perceptions of variation in a sampling situation. In T. Nakahara & M. Koyama (Eds.), Proceedings of the 24th Annual Meeting of the International Group for the Psychology of Mathematics Education , Hiroshima: Japan.

Rial, J. (1998). Randomness [Review of the book *Randomness*]. American Scientist, 86, 482.

Richards, L. (1995). Transition work! Reflections on a three-year NUD*IST Project. In R. Burgess (Ed.), Studies in Qualitative Methodology: Vol. 5. Computing and Qualitative Research. (pp. 105-140). Greenwich, CT: JAI Press.

Richards, T., & Richards, L. (1994). Using computers in qualitative research. In N. Denzin & Y. Lincoln (Eds.), Handbook of Qualitative Research (pp. 445-462). Thousand Oaks, CA: Sage Publications.

Roberts, R., & Scheaffer, R. (1999). Advanced Placement Statistics – Past, present, and future. American Statistician, 53 (4), 307-321.

Romberg, T. (1992). Perspectives on scholarship and research methods. In D. Grouws (Ed.), Handbook for Research on Mathematics Teaching and Learning , (pp. 49-64). New York: Macmillan.

Rubin, A., Bruce, B., & Tenney, Y. (1991). Learning about sampling: Trouble at the core of statistics. In D. Vere-Jones (Ed.), Proceedings of the Third International Conference on Teaching Statistics, Vol. 1 , (pp. 314-319). Voorburg, The Netherlands: International Statistical Institute.

Russell, S., Mokros, J., Goldsmith, L., & Weinberg, A. (1990). What's Typical? Teachers' Descriptions of Data. Paper presented at the Third International Conference on Teaching Statistics (ICOTS III), Dundedin, New Zealand.

Saldanha, L., & Thompson, P. (2001). Students' reasoning about sampling distributions and statistical inference. In C. Reading (Ed.), Background Readings for the Second International Research Forum on Statistical Reasoning, Thinking, and Literacy. Armidale: Australia.

Scheaffer, R. (2002). Data Analysis in the K-12 Mathematics Curriculum: Teaching the Teachers. Paper presented at a conference of the International Association of Statistics Education (IASE).

Seife, C. (1997). New test sizes up randomness. Science, 276, 532.

Shaughnessy, J. M. (1977). Misconceptions of probability: An experiment with a small-group, activity-based, model building approach to introductory probability at the college level. Educational Studies in Mathematics, 8 , 285-316.

Shaughnessy, J. M. (1992). Research in Probability and Statistics: Reflections and Directions. In D.A. Grouws (Ed.), Handbook of Research on Mathematics Teaching and Learning, (pp. 465-494). New York, NY: Macmillan.

Shaughnessy, J. (1993). Probability and statistics. Mathematics Teacher, 86 , 244-248.

Shaughnessy, J. (1997). Missed opportunities in research on the teaching and learning of data and chance. In F. Bidduch & K. Carr (Eds.), Proceedings of the 20th Annual Conference of the Mathematics Education Research Group of Australasia, (pp. 6-22). Rotorua, NZ: MERGA.

Shaughnessy, J. (2003).*The Development of Secondary Student's Conceptions of Variability.* Annual report year 1, NSF Grant No. REC 0207842. Portland, Oregon: Portland State University.

Shaughnessy, M. & Arcidiacono, M. (1993). Visual Encounters with Chance (Unit VIII, Math and the Mind's Eye). Salem, OR: The Math Learning Center.

Shaughnessy, J., & Bergman, B. (1993). Thinking about uncertainty: Probability and statistics. In P. Wilson (Ed.), Research Ideas for the Classroom: High School Mathematics, (pp. 177-197). New York: Macmillan.

Shaughnessy, J., & Ciancetta, M. (2001). Conflict Between Students' Personal Theories and Actual Data: The Spectre of Variation. Paper presented at the Second Roundtable Conference on Research on Statistics Teaching and Learning, Armidale, New South Wales, Australia.

Shaughnessy, J., & Ciancetta, M. (2002). Students' understanding of variability in a probability environment. Paper presented at The Sixth International Conference on Teaching Statistics, Cape Town, South Africa.

Shaughnessy, J., Ciancetta, M., & Canada, D. (2003). Middle school students' thinking about variability in repeated trials: A cross -task comparison. In N. Pateman, B. Dougherty, & J. Zillah (Eds.). *Proceedings of the 27th Conference of the International Group for the Psychology of Mathematics Education: Vol. 4*. Honolulu, HI: University of Hawaii.

Shaughnessy, J., Ciancetta, M., Best, K., & Canada, D. (2004). Students' Attention to Variability when Comparing Distributions. Paper Presented at the Research Presession of the 82nd Annual Meeting of the National Council of Teachers of Mathematics, Philadelphia, PA.

Shaughnessy, J., Garfield, J., & Greer, B. (1996). Data handling. In A. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), International Handbook of Mathematics Education (Part 1), (pp. 205-237). Dordrecht, The Netherlands: Kluwer.

Shaughnessy, J., & Pfannkuch, M. (2002). How faithful is Old Faithful? Statistical thinking: A story of variation and prediction. Mathematics Teacher, 95 (4) , 252-270.

Shaughnessy, J., Watson, J., Moritz, J., & Reading, C. (1999). School mathematics students' acknowledgement of statistical variation. In C. Maher (Chair), There's more to life than centers. Presession Research Symposium, 77th Annual National Council of Teachers of Mathematics Conference, San Francisco, CA.

Shrage, G. (1983). (Mis)-Interpretation of stochastic models. In R. Scholz (Ed.), Decision Making Under Uncertainty (pp. 351-361). Amsterdam, The Netherlands: North-Holland.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. Educational Researcher, 15, 4-14.

Shulman, L. (1988). Paradigms and Research Programs in the Study of Teaching: A Contemporary Perspective. In M.C. Whittrock (Ed.), Handbook of Research on Teaching, 3rd Edition (pp. 3- 35). New York: Macmillan.

Simon, M. (1993). Prospective elementary teachers' knowledge of division. Journal for Research in Mathematics Education, 24, 233-254.

Simon, M. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. Journal for Research in Mathematics Education, 26 (2), 146-159.

Stake, R. (1994). Case studies. In N. Denzin & Y. Lincoln (Eds.), Handbook of Qualitative Research (pp. 236-247). Thousand Oaks, CA: Sage Publications.

Stake, R. (1995). The Art of Case Study Research. Thousand Oaks, CA: Sage Publications.

Steinbring, H. (1990). The use of the chance-concept in everyday teaching – Aspects of a socially constituted epistemology of mathematical knowledge. In D. Vere-Jones (Ed.), Proceedings of the Third International Conference on Teaching Statistics , Vol. 1 , (pp. 1-24). Voorburg, The Netherlands: International Statistical Institute.

Strauss A., & Corbin, J. (1990). Basics of qualitative research: Grounded theory and procedures. London, UK: Sage Publications.

Strauss A., & Corbin, J. (1994). Grounded theory methodology: An overview. In N. Denzin & Y. Lincoln (Eds.), Handbook of Qualitative Research (pp. 273-285). Thousand Oaks, CA: Sage Publications.

Strauss, A., & Corbin, J. (1998). Basics of qualitative research (3rd ed.). Thousand Oaks, CA: Sage Publications.

Telese, J. (1999). The role of social constructivist philosophy in the teaching of school algebra and the preparation of mathematics teachers. Paper presented at the 79th Annual Meeting of the Association of Teacher Educators, Chicago, IL.

Teppo, A. (1997). Diverse ways of knowing. In A. Teppo (Ed.), Qualitative Research Methods in Mathematics Education (pp. 1-16). Reston, VA: National Council of Teachers of Mathematics.

Thompson, D. (Ed.). (1998). The Oxford Dictionary of Current English . New York, NY: Oxford University Press.

Tirosh, D. (2000). Enhancing prospective teachers' knowledge of children's conceptions: The case of division of fractions. Journal for Research in Mathematics Education, 31, 5-25.

Torok, R., & Watson, J. (2000). Development of the concept of statistical variation: An exploratory study. Mathematical Education Research Journal, 12 (2), 147-169.

Truran, J. (1994). Children's intuitive understanding of variance. In J. Garfield (Ed.), Research Papers from the 4th International Conference on Teaching Statistics (ICOTS 4). Minneapolis, MN: International Study Group for Research on Learning Probability and Statistics.

Truran, K. (1997). Beliefs about teaching stochastics held by primary pre-service teaching students. In F. Bidduch & K. Carr (Eds.), Proceedings of the 20th annual conference of the Mathematics Education Research Group of Australasia. Rotorua, NZ: MERGA.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. Psychological Bulletin, 76 (2), 105-110.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 1124-1131.

Von Glasersfeld, E. (1993). Questions and answers about radical constructivism. In K. Tobin, (Ed.), The Practice of Constructivism in Science Education (pp. 23-38). Hillsdale, NJ: Lawrence Erlbaum Associates.

Von Glasersfeld, E. (1995). Radical constructivism: A way of knowing and learning. London, UK: The Falmer Press.

Von Glasersfeld, E. (1996a). Aspects of radical constructivism and its educational recommendations. In L. Steffe, P. Nesher, P. Cobb, G. Goldin, & B. Greer (Eds.), Theories of Mathematical Learning (pp. 307-314). Mahwah, NJ: Lawrence Erlbaum Associates.

Von Glasersfeld, E. (1996b). Introduction: Aspects of constructivism. In C. Fosnot (Ed.), Constructivism: Theory, Perspectives, and Practice (pp. 3-7). New York, NY: Teachers College.

Von Glasersfeld, E. (1998). Why constructivism must be radical. In M. Larochelle, N. Bednarz, & J. Garrison (Eds.), Constructivism and Education (pp. 23-28). Cambridge, UK: Cambridge University Press.

Vygotsky, Lev. (1986). Thought and Language. (A. Kozulin, Trans.). Cambridge, MA: The MIT Press. (Original work published 1934).

Watson, J. (1997). Assessing statistical thinking using the media. In I. Gal & J. Garfield (Eds.), Handbook of Assessment in Statistics Education, (pp. 107-121). Amsterdam: IOS Press and the International Statistical Institute.

Watson, J. (1998). Assessment of statistical understanding in a media context. In L. Pereira-Mendoza, L.S. Kea, T.W. Kee, & W. Wong (Eds.), Proceedings of the Fifth International Conference on Teaching Statistics, Voorburg, The Netherlands: International Statistics Institute.

Watson, J. (2000a). Preservice mathematics teachers' understanding of sampling: Intuition or mathematics. Mathematics Teacher Education and Development, 2, 121-135.

Watson, J. (2000b). The development of school students' understanding of statistical variation. Australian Research Council (ARC) Project No. A00000716.

Watson, J. (2001). Profiling teachers' competence and confidence to teach particular mathematics topics: The case of chance and data. Journal of Mathematics Teacher Education 1 (2), 1-33.

Watson, J., Collis, K., & Moritz, J. (1995, November). The development of concepts associated with sampling in grades 3, 5, 7, and 9. Paper presented at the Annual Conference of the Australian Association for Research in Education. Hobart, Tasmania.

Watson, J., Kelly, B., Callingham, R., & Shaughnessy, J. (2002). The measurement of school students' understanding of statistical variation. The International Journal of Mathematical Education in Science and Technology. (34), 1-29.

Watson, J., & Moritz, J. (1997). Teachers' views of sampling. In N. Scott & H. Hollingsworth (Eds.) Mathematics Creating the Future, (pp. 345-353). Adelaide: Australian Association of Mathematics Teachers, Inc.

Watson, J., & Moritz, J. (1999). The beginning of statistical inference: Comparing two data sets. Educational Studies in Mathematics, 37 (2), 145-168.

Watson, J., & Moritz, J. (2000a). The development of concepts associated with sampling. Journal of Mathematics Behavior.

Watson, J., & Moritz, J. (2000b). Developing Concepts of Sampling. Journal for Research in Mathematics Education, 31 (1), 44-70.

Watson, J., & Moritz, J. (2000c). The longitudinal development of understanding of average. Mathematical Thinking and Learning, 2, 11-50.

Well, A., Pollatsek, A., & Boyce, S. (1990). Understanding the effects of sample size on the variability of the mean. Organizational Behavior and Human Decision Processes, 47, 289-312.

Wild, C., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. International Statistical Review, 67 , 233-265.

Zawojewski, J., & Shaughnessy, J. (2000). Data and chance. In E. Silver & P. Kenney (Eds.), Results from the seventh mathematics assessment of the National Assessment of Educational Progress (pp. 235-268). Reston, VA: National Council of Teachers of Mathematics.

APPENDIX A

Informed Consent

You are invited to participate in a doctoral research project entitled "Elementary Preservice Teachers' Conceptions of Variation", being conducted by Daniel Canada from the Department of Mathematical Sciences at Portland State University. The researcher hopes to develop a characterization of the knowledge held by elementary preservice teachers about this important statistical concept. You were selected as a possible participant by virtue of your enrollment in the Math 212 class.

By giving your consent to take part in this study, you are agreeing to three distinct aspects of data gathering. First of all, comments made by you in class which the researcher deems pertinent can be transcribed and used as data. Secondly, homework which is relevant to the project can be photocopied and used as data. Thirdly, you agree to participate in at least one interview which takes place outside of the normal class hours. The interview will be scheduled at a mutually convenient time and place; it will be videotaped, and will last approximately one hour. The transcripts from this interview can also be used as data.

You as a prospective teacher will gain a direct benefit from a deeper exploration of your own ideas about this key statistical concept; this exploration allows you to extend your own learning about variation in the non-evaluative environment of the research project. Moreover, the practice in articulating your thinking is especially helpful as you make the transition to your own classroom, and invoke similar practices with your own students.

Potential risks include the possibility that an unauthorized person may view the data, or that your actual name may inadvertently become associated with the data. To minimize this risk, all written responses, notes, audio and video tapes, and transcriptions will be kept confidential, and will be kept locked up in the researcher's office in the Department of Mathematical Science at PSU. After three years, these records will be destroyed. In writing any results for the study, pseudonyms will be used so that your identity cannot be matched with the responses you have provided. There is also a risk that having a researcher in the classroom may affect the learning environment. To lessen this risk, it will be stressed that this research is descriptive and not evaluative in nature.

Your participation in this study is voluntary and you are completely free to withdraw from the study at any time. Your decision to participate or not will not affect your relationship with the researcher or with any academic program at PSU in any way.

If you have concerns about your participation in this study or your rights as a research subject, please contact the Human Subjects Research Review Committee, Office of Research and Sponsored Projects, 111 Cramer Hall, Portland State University, (503) 725-8182. If you have any questions about the study itself, please contact Daniel Canada, at the Department of Mathematical Sciences, 334 Neuberger Hall, Portland State University, (503) 725-3621.

Your signature indicates that you have read and understand the above information and agree to take part in this study. Please remember that you may withdraw your consent at any time without penalty. Also, by signing, you are not waiving any legal claims, rights or remedies. The researcher has provided you with a copy of this form for your records.

_____ _____

Signature of Participant Date

_____ _____

Daniel Canada, Researcher Date
Department of Mathematical Sciences
Portland State University
(503) 725-3621

APPENDIX B

Surveys and Interviews

The surveys and interview scripts are appended in the order that they were administered:

- PreSurvey p. 327
- PreInterview p. 336
- PostSurvey (Data & Graphs) p. 348
- PostSurvey (Sampling) p. 352
- PostSurvey (Probability) p. 355
- PostInterview p. 357

PRESURVEY

Mth 212 Survey on Probability and Statistics Name:

1] Where & when did you take Math 211?

2] If you have taken prior math courses in which probability and/or statistics was taught,

a) When & where did you take those courses ?

b) How did you feel about the probability and/or statistics at that time?

3] How comfortable do you feel about learning probability and/or statistics now ?

4] What does the word "random" mean to you ?

Give an example of something that happens in "a "random" way.

5] What does the word "variation" mean to you ?

Give an example of something that "varies".

This set of questions helps to give a picture of how you think about some problems in probability and statistics. Rather than think in terms of a right or wrong answer, just write down your best thinking for each situation. Later in the quarter, we'll explore situations like these as a class.

[1] Suppose there is a container with 100 pieces of candy in it. 60 are Red, and 40 are
Yellow. The candies are all mixed up in the container.

You reach in and pull out a handful of 10 candies at random.

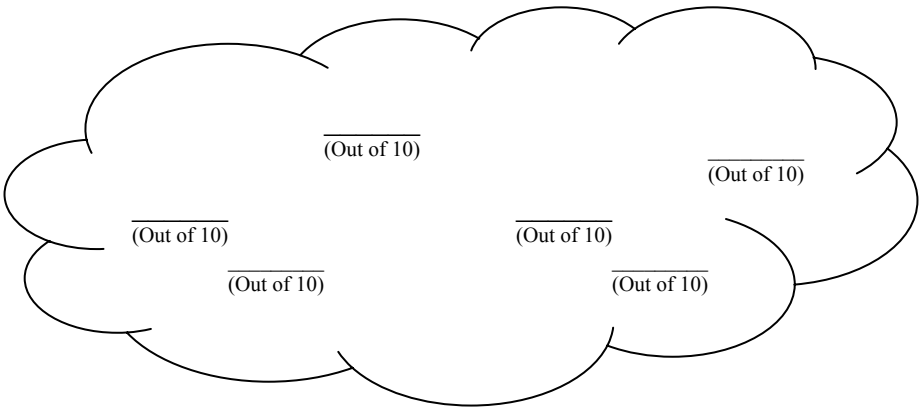
(a) How many red candies do you think you might get?

Why do you think this?

(b) Suppose you do this several times (each time returning the previous handful of 10 candies and remixing the container). Do you think this many reds would come out every time?

Why do you think this?

(c) Suppose six classmates do this experiment (each time returning the previous handful of 10 candies and remixing the container). Write down the number of reds that you think each classmate might get:



Why did you choose those numbers?

[2] Suppose 6 people did this experiment – pulled ten candies from the container, wrote down the number of reds, then returned the ten and remixed all the candies.

What do you think the numbers of reds will most likely go from?

From a low of _____ to a high of _____.

Now suppose 30 people did this experiment. What do you think the numbers of reds will most likely go from?

From a low of _____ to a high of _____.

Why do you think this?

[3] At the same container, suppose that 50 people each pulled out handfuls of 10 candies, wrote down the number of reds, put the candies back and mixed them up again. Of the 50 people, how many of them do you think would get:

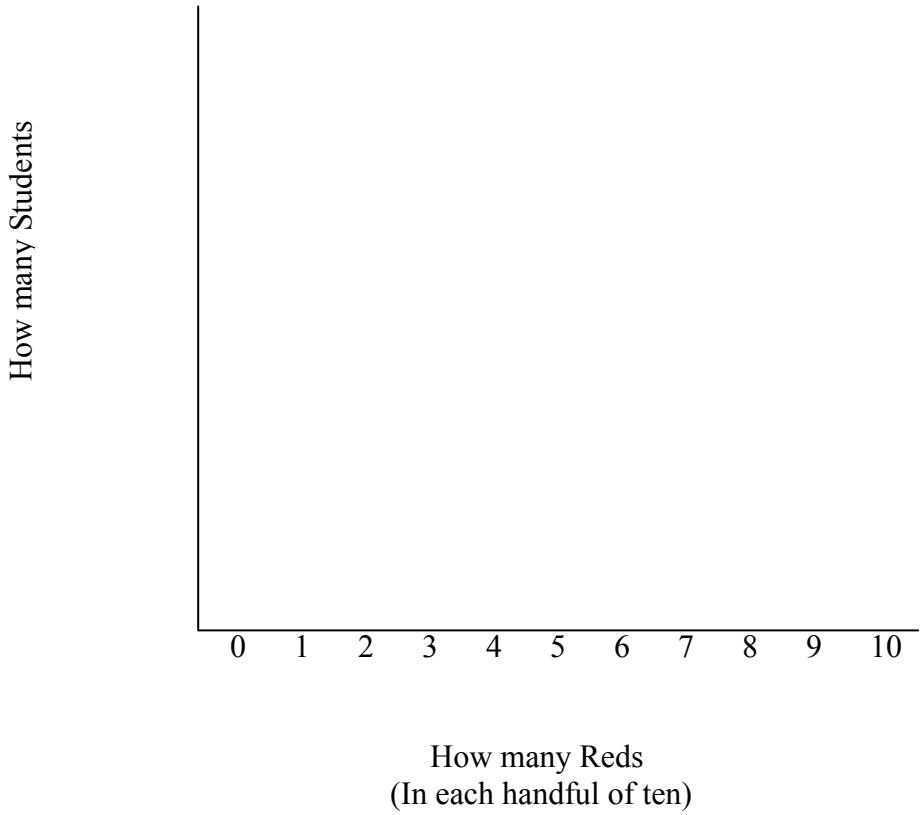
0 Red ? ____
1 Red ? ____
2 Red ? ____
3 Red ? ____
4 Red ? ____
5 Red ? ____
6 Red ? ____
7 Red ? ____
8 Red ? ____
9 Red ? ____
10 Red ? ____

Total : 50 People

Why do you think the numbers you wrote above are reasonable?

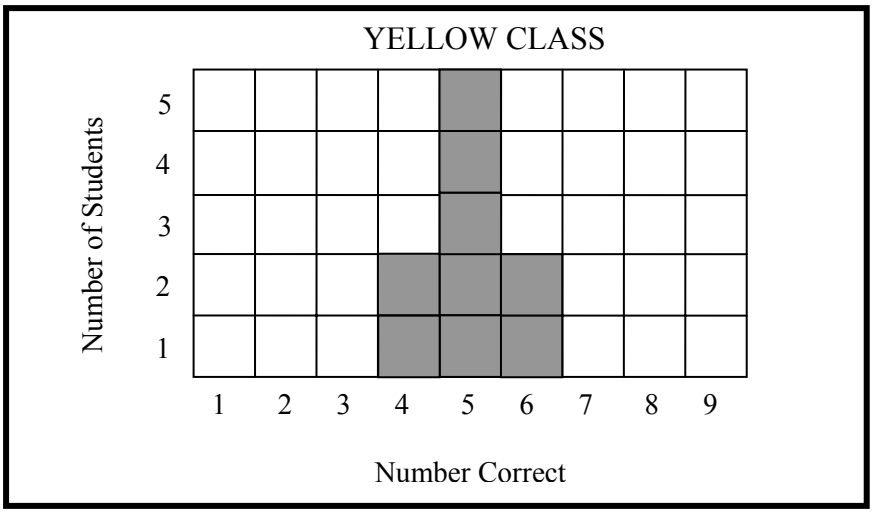
[4] Fifty students lined up at the candy container. Each student pulled a handful of 10 candies, wrote on the chalkboard how many reds they had, and then returned the candies and mixed them all up again.

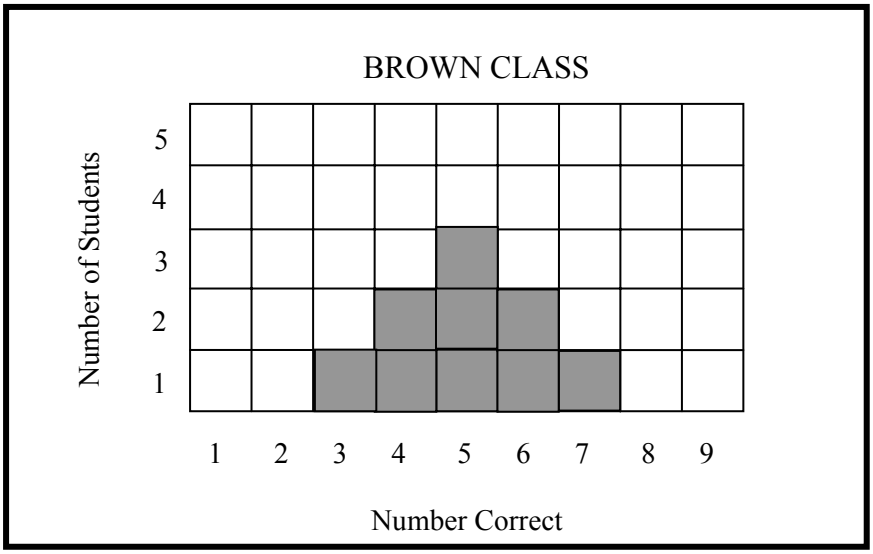
The class decided to draw a graph of their data. Show below what their graph might look like:



[5] Two schools are comparing some classes to see which school is better at spelling. All the classes took identical tests.

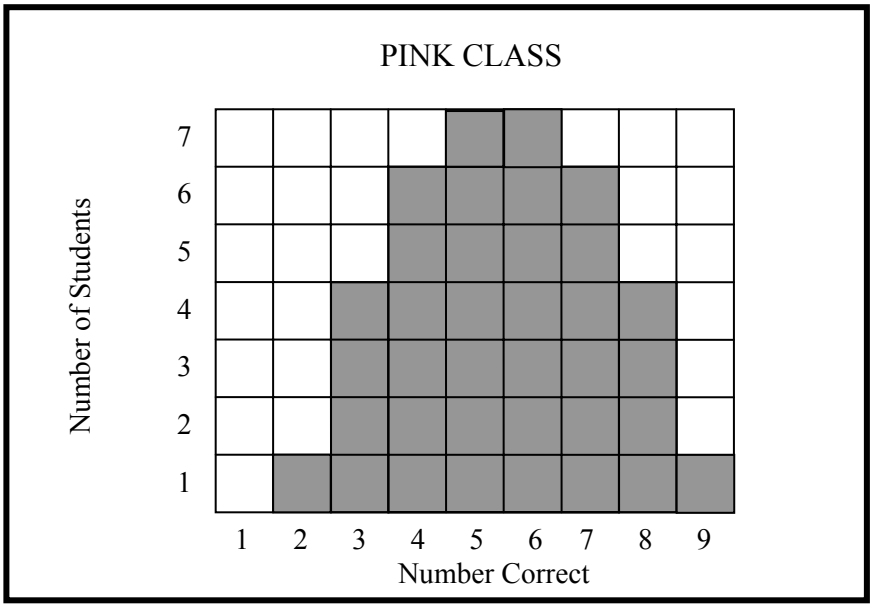
(a) First consider two classes, the YELLOW class and the BROWN class. The scores for the two classes are shown on the two charts below, Each shaded box is one person's test score.

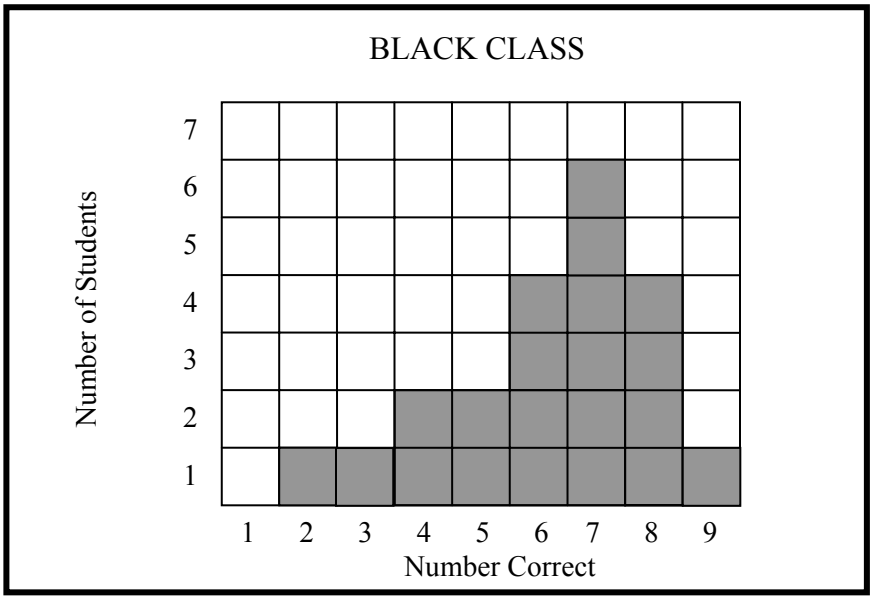




Now look at the scores of all students in each class, and then decide: Did the two classes do equally well on the test, or did one of the classes do better than the other? Explain how you decided.

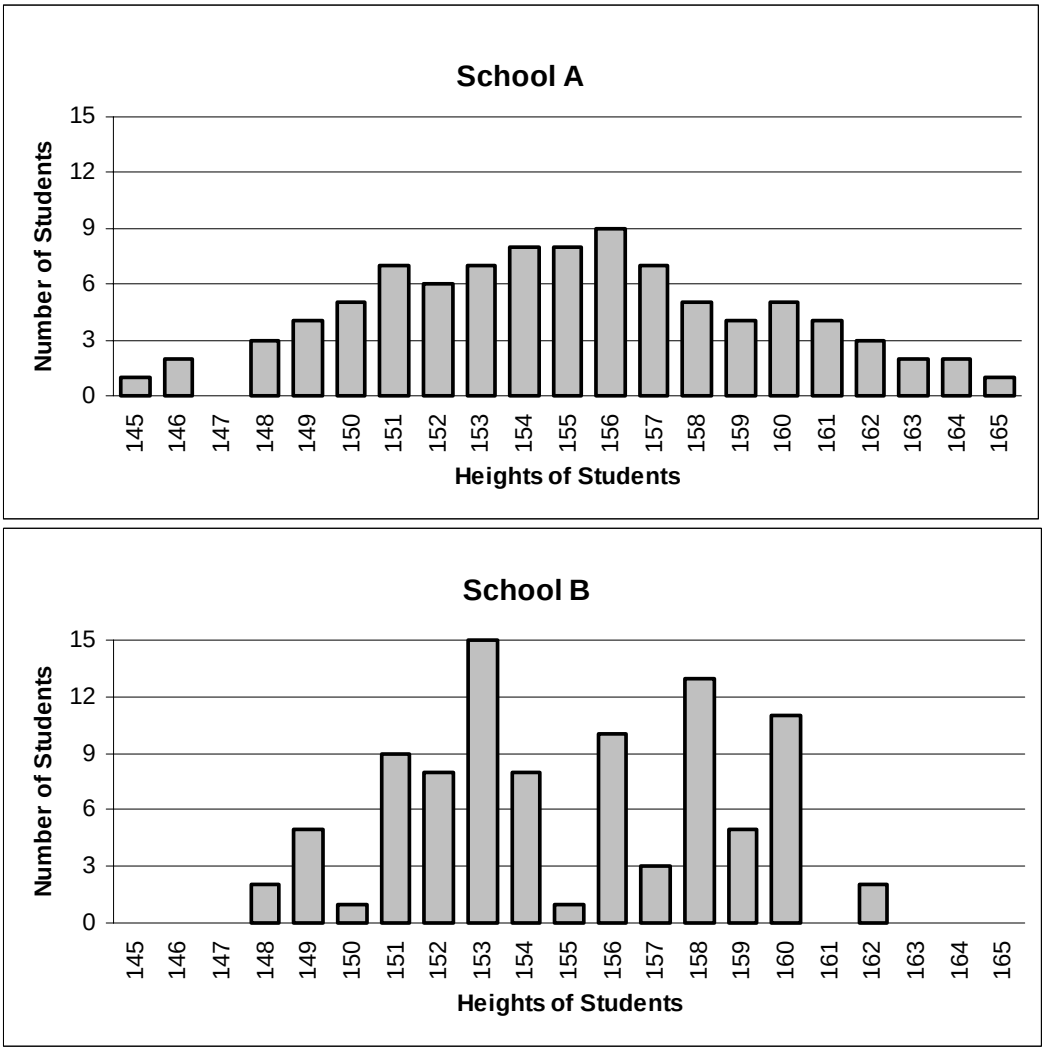
(b) Now consider two more classes, the PINK class and the BLACK class. The scores for the two classes are shown below, and once again each box is one person's test score.





Again look at the scores of all students in each class, and then decide: Did the two classes do equally well on the test, or did one of the classes do better than the other? Explain how you decided.

[6] The following graphs describe some data collected about Grade 7 students' heights (measured in centimeters) in two different schools:



Which graph shows more variability in students' heights?

Explain why you think this.

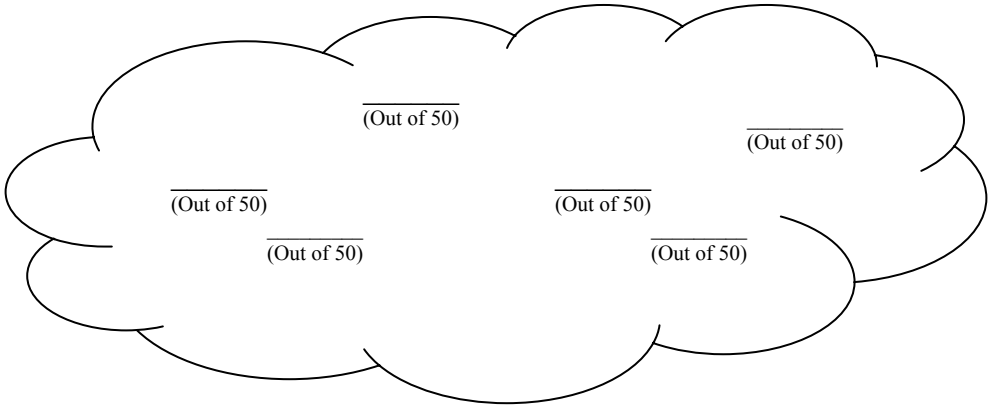
[7] Consider flipping a fair coin.

(a) Mark is curious to see how often the coin lands Heads-up, so he flips it 50 times. How many times out of 50 flips do you think the coin might land Heads-up for Mark?

Why do you think this?

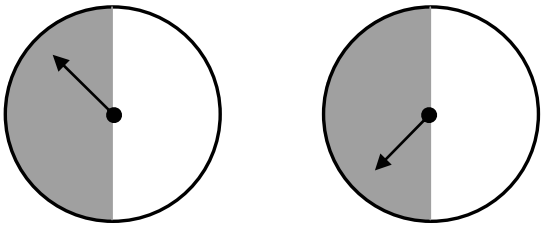
(b) After Mark's first set of 50 flip, he decides to do a second set of 50 flips. How do you think his results on the second set of 50 flips will compare with the results of his first set?

(c) Mark actually has a lot of time on his hands, so the next day he does 6 sets of 50 flips. Write in the numbers for what you think might happen for the number of flips out of 50 the coin would land Heads-up (in each of the 6 sets of 50 flips).



Why did you choose those numbers?

[8] The two fair spinners shown below are a part of a game, which goes like this: A player spins each spinner once, and wins a prize *only if both* arrows land on **black**.

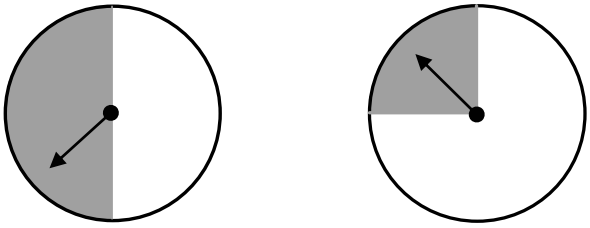


Angela thinks she has a 50-50 chance of winning. Do you agree?

☐ Yes ☐ No

Explain your answer. Why do you think this?

[9] Suppose the game is played with the new spinners shown below. Again, both spinners are spun once, and the arrows must *both* land on **black** in order to win.



What do you think the chances of winning this game would be?

Explain your answer. Why do you think this?

PREINTERVIEW

First Interview Date: Name:

[1] Suppose there is a container with 100 pieces of candy in it. 60 are Red, and 40 are Yellow. The candies are all mixed up in the container.

You reach in and pull out a handful of 10 candies at random.

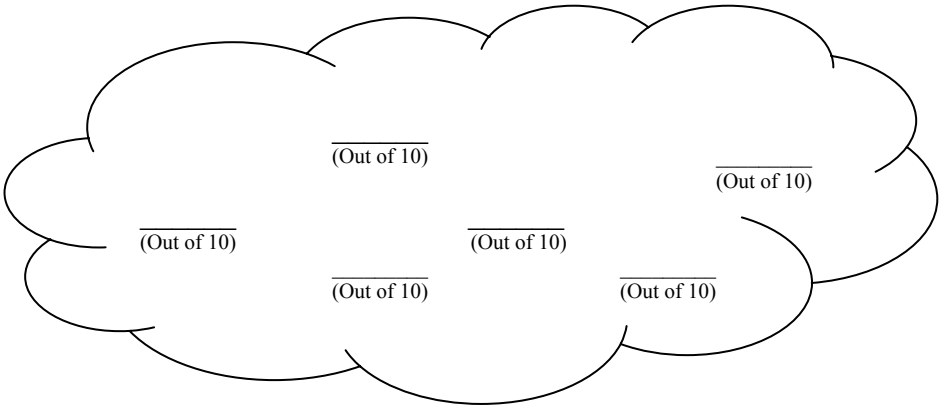
(a) How many red candies do you think you might get?

Why do you think this?

(b) Suppose you do this several times (each time returning the previous handful of 10 candies and remixing the container). Do you think this many reds would come out every time?

Why do you think this?

(c) Suppose six classmates do this experiment (each time returning the previous handful of 10 candies and remixing the container). Write down the number of reds that you think each classmate might get:



Why did you choose those numbers?

[2] Here are some examples of what other people have said for the numbers of reds that they think the six classmates would get in each handful.

i) 7, 9, 7, 6, 8, 7

ii) 6, 7, 5, 8, 5, 4

iii) 6, 6, 6, 6, 6, 6

iv) 2, 5, 4, 3, 6, 4

v) 3, 10, 9, 2, 1, 5

(a) Put a check mark next to any of these that you think might be likely.

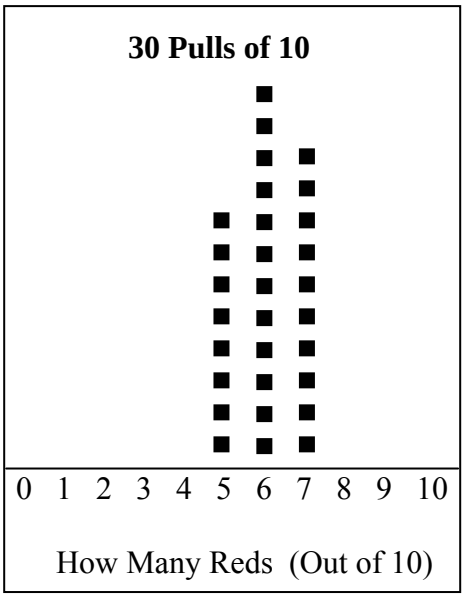
(b) Circle the list that you think best describes what might happen.

(c) Why do think the list you chose best describes what might happen?

[3] Matt took his class to the candy container (100 Candies = 60 Red and 40 Yellow). Then he left the room. When he came back, the class claimed to have pulled 30 samples each of size 10, with replacement. They showed Matt their data and a graph:

Number of Reds
in 30 Samples of 10

7	6	5
5	7	6
6	5	7
7	6	7
6	6	6
7	5	5
5	6	5
6	7	6
7	6	6
5	7	7



Which of the following do you think is *most* likely ? Put a check mark next to it.

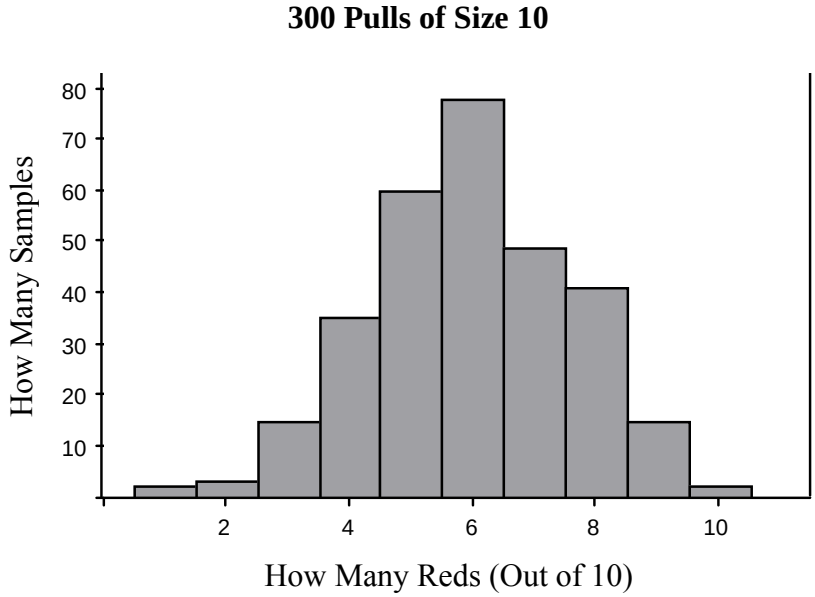
___ Matt's class just made up these results

___ Those are the actual results of the class samples

___ No one can have much confidence if the results are made up or not.

Explain why you think this is the most likely.

[4] Jen's class also visited the candy container (100 Candies = 60 Red and 40 Yellow). The class claims to have pulled 300 samples each of size 10, with replacement. They showed Jen this graph:



(a) Which of the following do you think is *most* likely ? Put a check mark next to it.

___ Jen's class just made up these results

___ Those are the actual results of the class samples

___ No one can have much confidence if the results are made up or not.

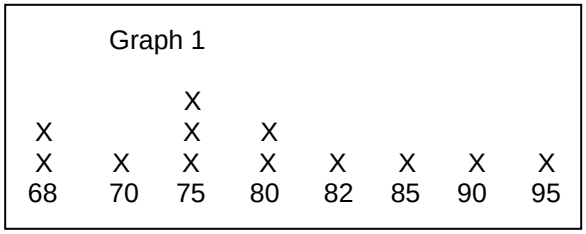
Explain why you think this is the most likely.

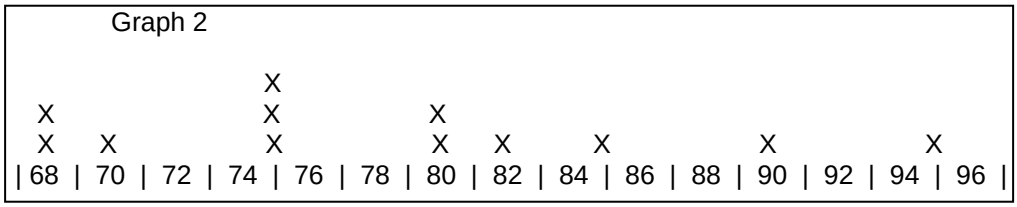
(b) How does the shape of the graph for Jen's class compare to the shape of the graph for Matt's class?

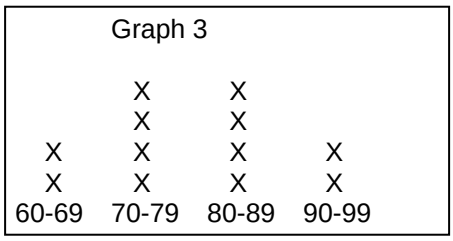
[5] A new car was being tested to see how well the brakes worked. The test engineer measured how many inches the car took to slow from 40 mph to 0 mph; the fewer inches taken, the better the braking power. Twelve trials were run, under the same road conditions and with the same test driver. Here were the results (to the nearest inch):

Stopping Distance (in.)			
68	68	70	75
75	75	80	80
82	85	90	95

The engineer was then trying to decide how to graph the results. She came up with the following three graphs for representing the data:







(a) Do these graphs differ in the way they show the braking power? If so, how?

(b) Do you think one graph shows more variability in the results than the others? Explain.

(c) If the engineer wanted to suggest that the car was fairly consistent in its braking power, which graph would you suggest she use, and why?

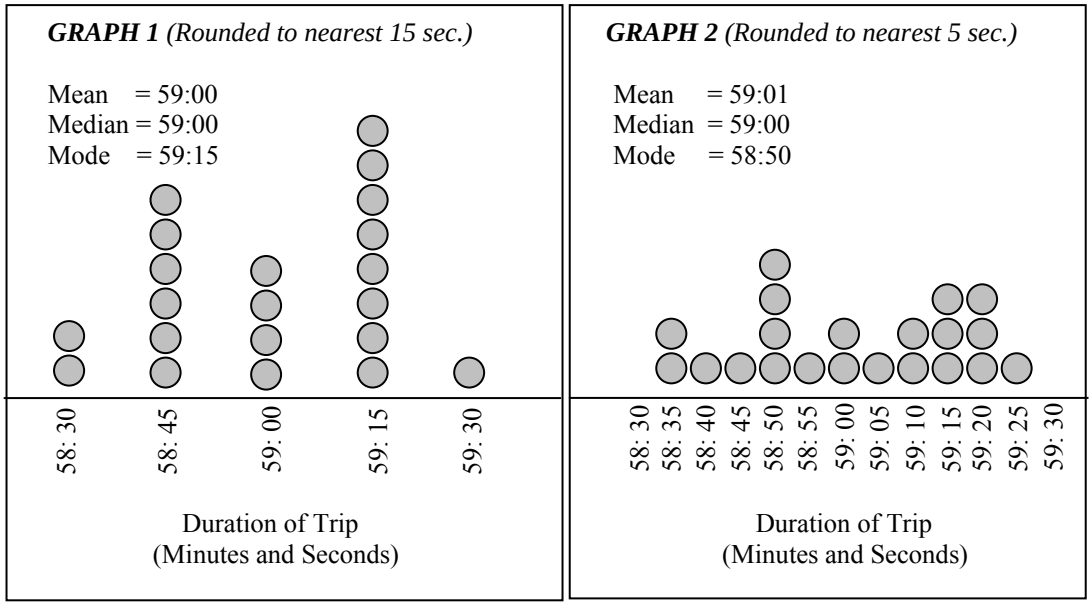
[6] A class of twenty-one 6th-grade students wanted to find out some information about MAX train rides. Their first goal was to find out the duration of a ride from Washington Park to Gresham. They all got on the same train, but they sat separately and kept track of the time on their own. Later in class, they were surprised to find that they did not have the same results:

Duration of Ride (Min:Sec , to the nearest second)

58: 36	58 :36	58: 40	58: 44	58: 51
58: 50	58: 49	58: 50	58 :56	59: 01
59: 02	59: 06	59: 11	59: 09	59: 16
59 :14	59: 15	59: 19	59: 21	59: 20
59: 24				

What are some possible reasons for why the class did not all get the same result?

[7] The class was deciding how to display their data. In Graph 1, they rounded to the nearest 15 seconds. In Graph 2, they rounded to the nearest 5 seconds.



(a) How do these graphs differ in the stories they tell about the duration of the trip?

(b) Some members of the class argue that the trip was really under 59 minutes, while some argue that it was over 59 minutes. Others claim it was exactly 59 minutes. What do you think about the true duration of the trip, and why do you think this?

(c) Does one graph help you more than the other in making your conclusion?

[8] The Wait-Time for the MAX is defined as the interval of time which starts when one train leaves and ends when the next train arrives. In other words, the Wait-Time is how long there's no train at the station.

A class of twenty students wanted to find out if there was a difference in Wait-Times between Westbound and Eastbound MAX trains. They went and got the following ten Wait-Times for different Westbound trains and ten Wait-Times for different Eastbound trains (rounded to the nearest half-minute):

Data: (Wait-Times in Minutes)

Westbound

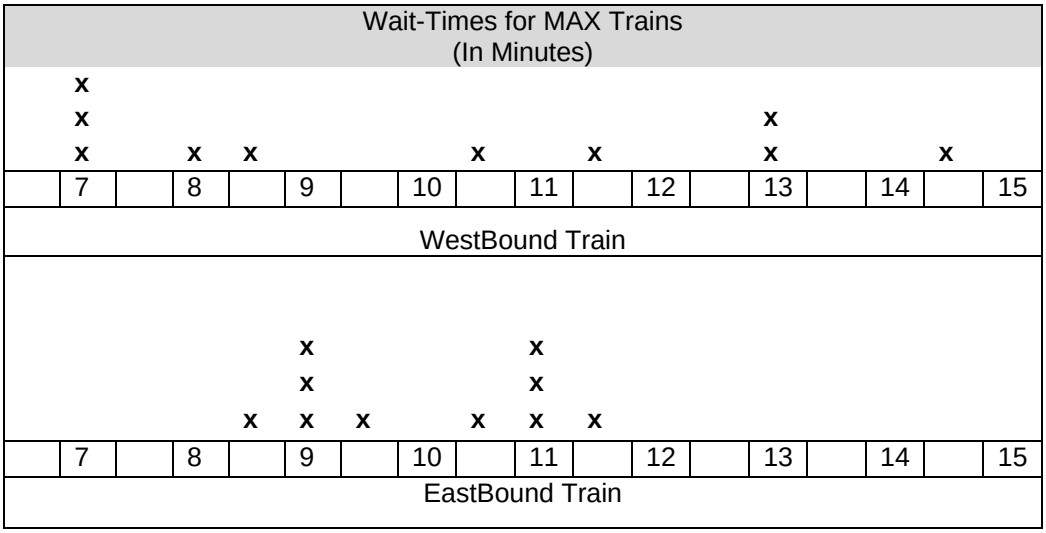
7.0	7.0	7.0	11.5	10.5
8.5	8.0	13.0	14.5.	13.0

Mean = 10.0 min.
Median = 10.5 min.

Eastbound

8.5	9.0	9.0	11.0	11.0
9.5	9.0	11.0	10.5	11.5

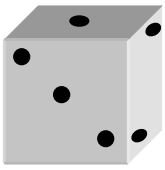
Mean = 10.0 min.
Median = 10.5 min.



(a) What can you conclude about the Wait-Times for the two trains?

(b) One student in class argues that there's really no difference in the Wait-Times of the two trains, since the averages are the same. Do you agree?

[9] Consider a regular, fair, six-sided die. Imagine that you threw the die 60 times. Fill in the table below to show how many times you think each number might come up.



Number that shows on the tossed die	How many times it might come up
1	
2	
3	
4	
5	
6	
Total =	**60**

Why do you think those numbers are reasonable?

[10] For homework, Mr. Blair asked each student in his class to toss a die 60 times and keep track of how many times each of the 6 sides came up. Below are the results turned in the next day by four students (Riki, for example, reported that Side 1 came up 7 times in 60 tosses).

	Riki	**Lynn**	**Lee**	**Pat**
Side that came up				
1	7	10	10	2
2	12	11	10	15
3	6	10	10	10
4	9	10	10	28
5	14	9	10	1
6	12	10	10	4

Only one of these students actually rolled the die. The other three students just made up their results before class. What do you think is most likely?

___i) Riki really rolled it

___ii) Lynn really rolled it

___iii) Lee really rolled it

___iv) Pat really rolled it

___v) No one can say. Any of the 4 students is equally likely to have really rolled it.

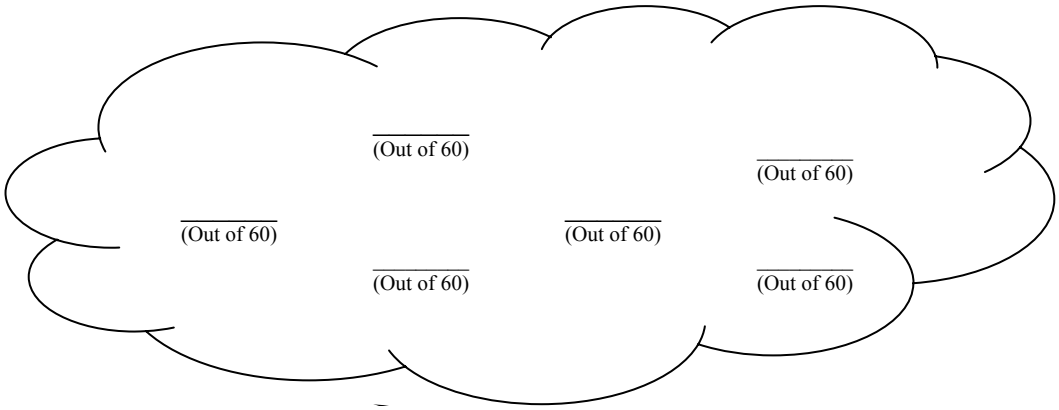
Explain your reasoning.

[11] Look back at Question [9] to see how many "5"s you predicted in 60 tosses.

(a) If you did another set of 60 tosses, do you think you would get that many "5"s again?

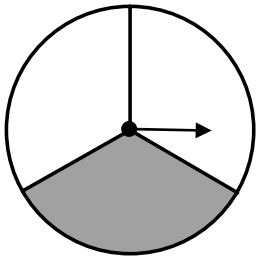
Why or why not?

(b) If six friends took turns tossing the die 60 times each, write down how many "5"s you think each friend might get in their 60 tosses:



Why did you choose those numbers?

[12] The spinner at the right has three regions of equal area: Two of the three regions are White and one of the three regions is Black.



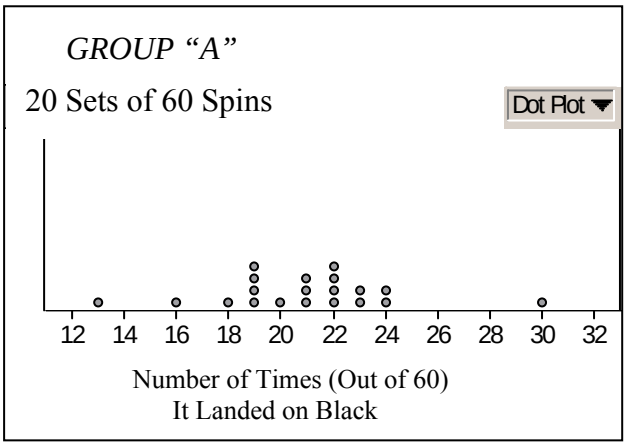
(a) If you spun this 3 times, would you be surprised if you got more Black than White?

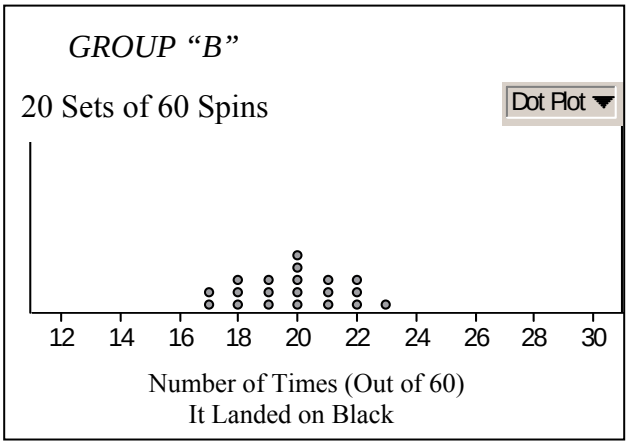
(b) If you spun this 12 times, would you be surprised if you got more Black than White?

(c) If you spun this 60 times, would you be surprised if it landed on White 30 times?

[13] Ron split his class into two groups, and he told each group to conduct the following experiment twenty times: Spin the spinner 60 times, and write down how many times it landed on Black. So, each group was supposed to do twenty sets of 60 spins.

Ron went for some coffee, and when he returned he saw these two graphs:





(a) What are some similarities and differences you notice in the two graphs above?

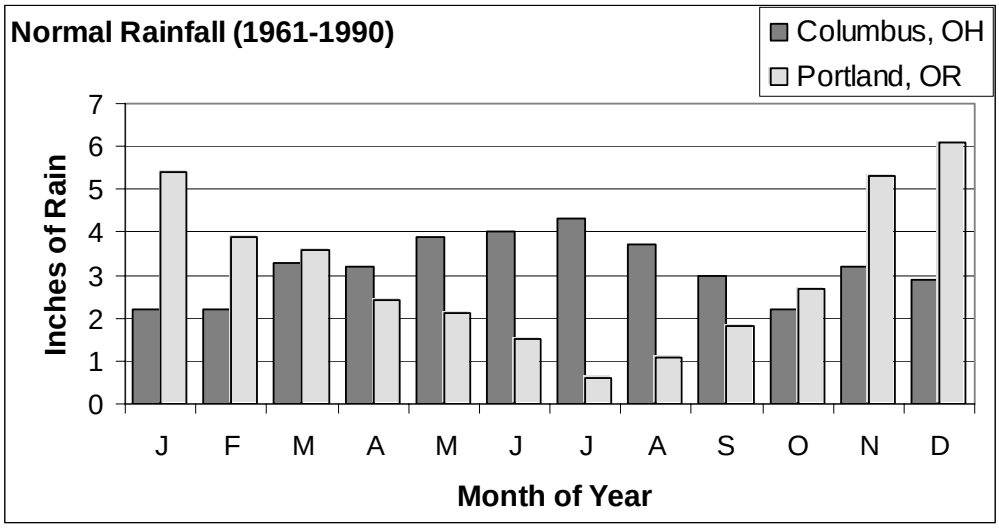
(b) Does one graph or the other look more like what you would have expected? Explain.

POSTSURVEY (DATA & GRAPHS)

Mth 212 Written Reflection: Data & Graphs Name: __

[1] The following data was taken from a National Weather Service. They kept records of the rainfall in cities to see how much rain fell each month. After 30 years, they averaged the amounts of rainfall in each month: This is called the average, or Normal Rainfall for the 30-year period.

(a) In the bar chart below, the normal monthly rainfall data for both Portland (Oregon) and Columbus (Ohio) are graphed together:

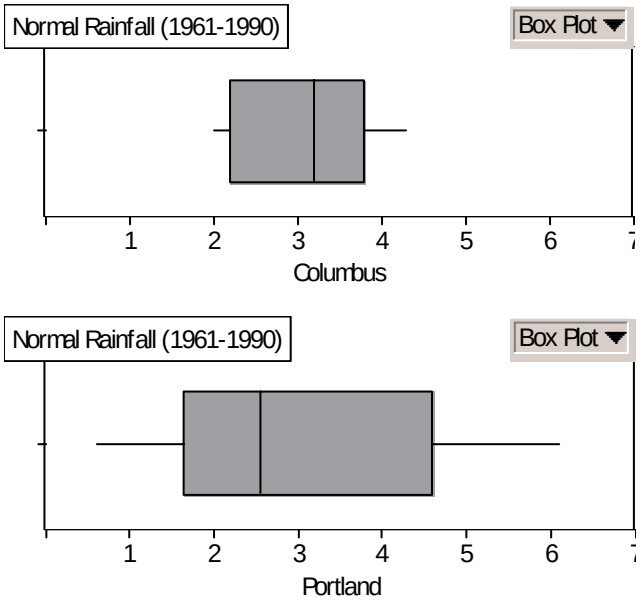


Columbus	Portland
Mean = 3.18 in.	Mean = 3.04 in.
Median = 3.20 in.	Median = 2.55 in.

i) What do you think are some causes for the different patterns of rain in the two cities?

ii) Adam and Zain are two Math 212 students who were discussing the data. Adam said Portland was rainier because it got the highest amount of rain a month. What do you think Adam is thinking when he says this?

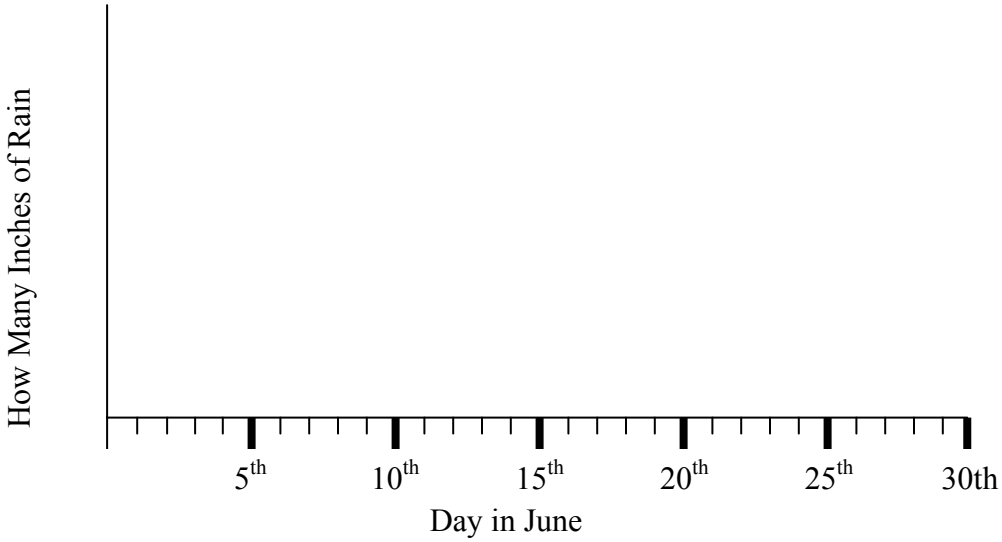
(b) Here are two boxplots that show the same data for the normal monthly rainfall in both Columbus and Portland:



i) Zain said Columbus was rainier because the average monthly rainfall was higher than Portland. What do you think Zain is thinking when he says this?

ii) Which city do *you* think is rainier, and why?

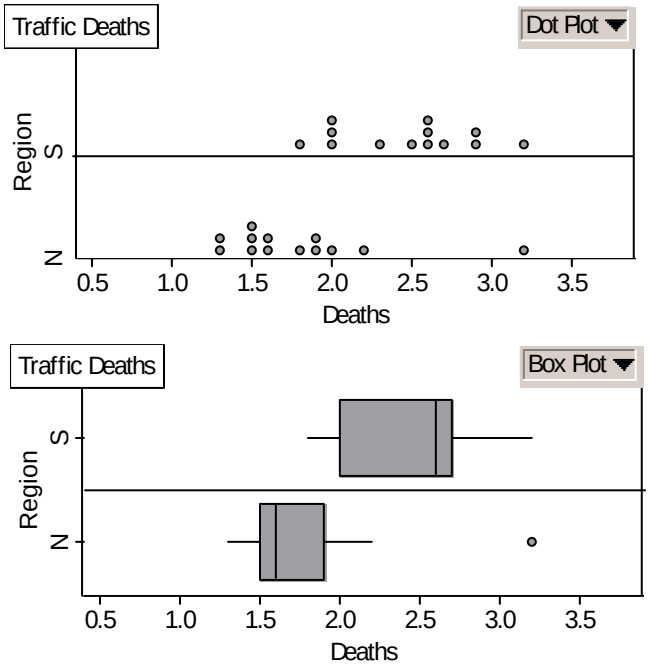
c) In Columbus, the normal monthly rainfall for the month of June is reported as 4 inches. Draw a graph below which shows how many inches of rain Columbus might get for each day in June (assuming that the average rainfall for the entire month is 4 inches).



[2] The National Safety Council groups the following 25 States plus the District of Columbia into two regions: South and Northeast. The following data and graphs show the number of traffic deaths in a recent year per 100 million vehicle miles driven:

Motor Vehicle Traffic Deaths per 100 Million Vehicle Miles Driven

South		**Northeast**	
Alabama	2.6	Connecticut	1.5
Arkansas	2.9	Delaware	2.2
Florida	2.7	District of Columbia	1.6
Georgia	2.0	Maine	1.8
Kentucky	2.6	Maryland	1.9
Louisiana	2.5	Massachusetts	1.3
Mississippi	3.2	New Hampshire	1.6
North Carolina	2.3	New Jersey	1.5
Oklahoma	2.0	New York	2.0
South Carolina	2.9	Pennsylvania	1.9
Tennessee	2.6	Rhode Island	1.3
Texas	2.0	Vermont	1.5
Virginia	1.8	West Virginia	3.2



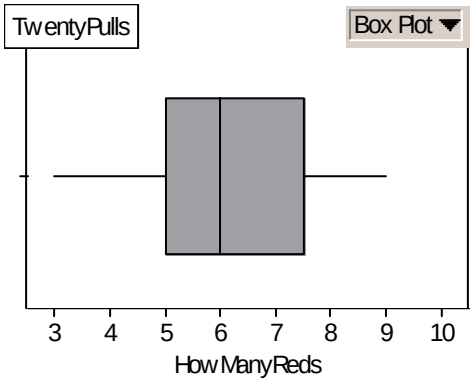
(a) How do the traffic deaths rates in the South compare with those in the Northeast ?

(b) What factors do you think might help to explain the difference between the South and the Northeast ?

POSTSURVEY (SAMPLING)

Mth 212 Written Reflection: Sampling Name: _____

[1] Consider the container of candy that holds 100 pieces (60 Red and 40 Yellow). Suppose that 20 people lined up at the container. Each person pulled out a handful of 10 candies at a time , wrote down the number of reds, put the candies back and mixed them up again. They made a boxplot of their results, and the graph looked like this:



Median = 6
Minimum = 3
Maximum = 9

a) When we look at the boxplot, we only get a bit of information about the entire data set. Write down what you think all 20 results might have been (for the number of red in each handful of ten):

____ ____ ____ ____ ____ ____ ____ ____ ____ ____

____ ____ ____ ____ ____ ____ ____ ____ ____ ____

b) Mike was surprised that nobody got 0 or 1 Red candy in their handful. He decides that he's going to try it with more than 20 people.

How many people do you think Mike should have do this so that at least one person gets 0 or 1 Red candy in their handful ?

How did you decide on that answer?

[2] Now suppose there is a LARGER container with 1000 pieces of candy in it. 600 are Red, 400 are Yellow. The candies are all mixed up in the container.

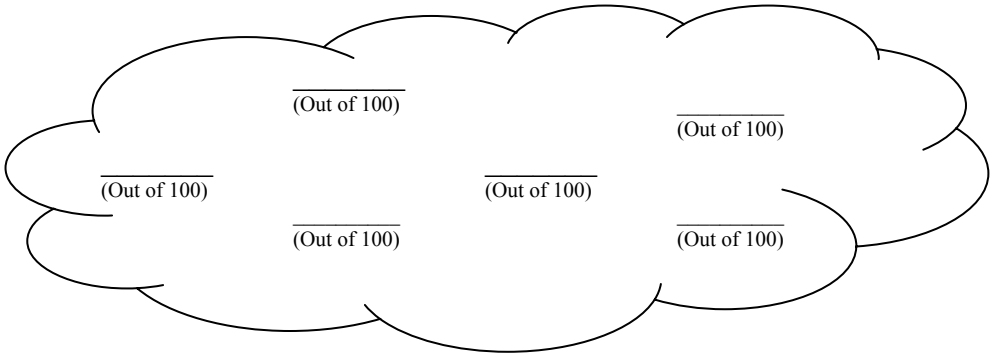
You reach in and pull out a handful of 100 candies at random.

(a) How many red candies do you think you will get?

(b) Suppose you do this several times (each time returning the previous handful of 100 candies and remixing the container). Do you think this many reds would
come out every time?

Why do you think this?

(c) Suppose six classmates do this experiment (each time returning the previous handful of 100 candies and remixing the container). Write down the number of reds that you think each classmate obtained:



Why did you choose those numbers?

[3] Suppose 30 people did this experiment – pulled one hundred, candies from the LARGE container (600 Red and 400 Yellow), wrote down the number of reds, then returned the hundred and remixed all the candies.

What do you think the numbers of reds will most likely go from?

From a low of _____ (out of 100) to a high of _____ (out of 100).

Now suppose 300 people did this experiment. What do you think the numbers of reds will most likely go from?

From a low of _____ (out of 100) to a high of _____ (out of 100).

Why do you think this?

[4] Suppose that 50 people each pulled out 100 candies from the LARGE container (600 Red and 400 Yellow), wrote down the number of reds they pulled, put the candies back and mixed them up again. Of the 50 people, how many of them do you think would get:

0 -10 Red ? ____
11-20 Red ? ____
21-30 Red ? ____
31-40 Red ? ____
41-50 Red ? ____
51-60 Red ? ____
61-70 Red ? ____
71-80 Red ? ____
81-90 Red ? ____
91-100 Red ? ____

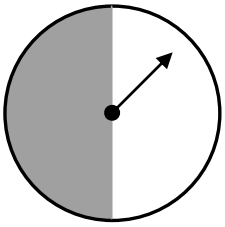
Total : 50 People

Why do you think the numbers you wrote above are reasonable?

POSTSURVEY (PROBABILITY)

Mth 212 Written Reflection: Probability Name: _____

[1] Consider the spinner on the right:

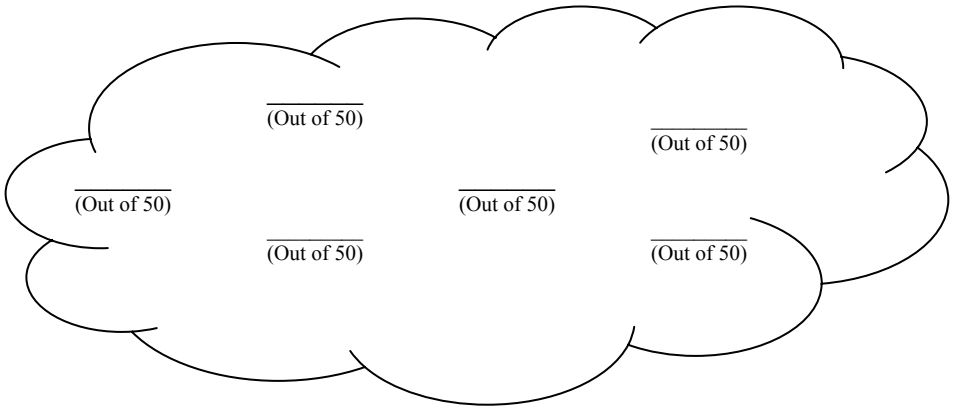


(a) Matt is curious to see how often the spinner lands on black, so he spins it 50 times. How many times (out of 50 tries) do you think the arrow might land black?

Why do you think this?

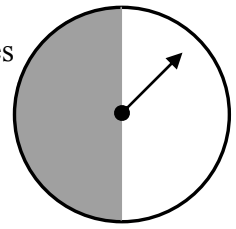
(b) After Matt's first set of 50 spins, he decides to do a second set of 50 spins. How do you think his results on the second set of 50 spins will compare with the results of his first set?

(c) Matt actually has a lot of time on his hands, so the next day he does 6 sets of 50 spins. Write a list that would describe what you think might happen for the number of spins out of 50 the spinner would land on black in each of the 6 sets of 50 spins.



Why did you choose those numbers?

[2] Suppose 30 students did this experiment – Each student spun the spinner 50 times and wrote down how many times (out of 50 tries) the arrow landed on black.



What do you think the students' numbers (of times the arrow lands on black) will most likely go from?

From a low of _____ to a high of _____.

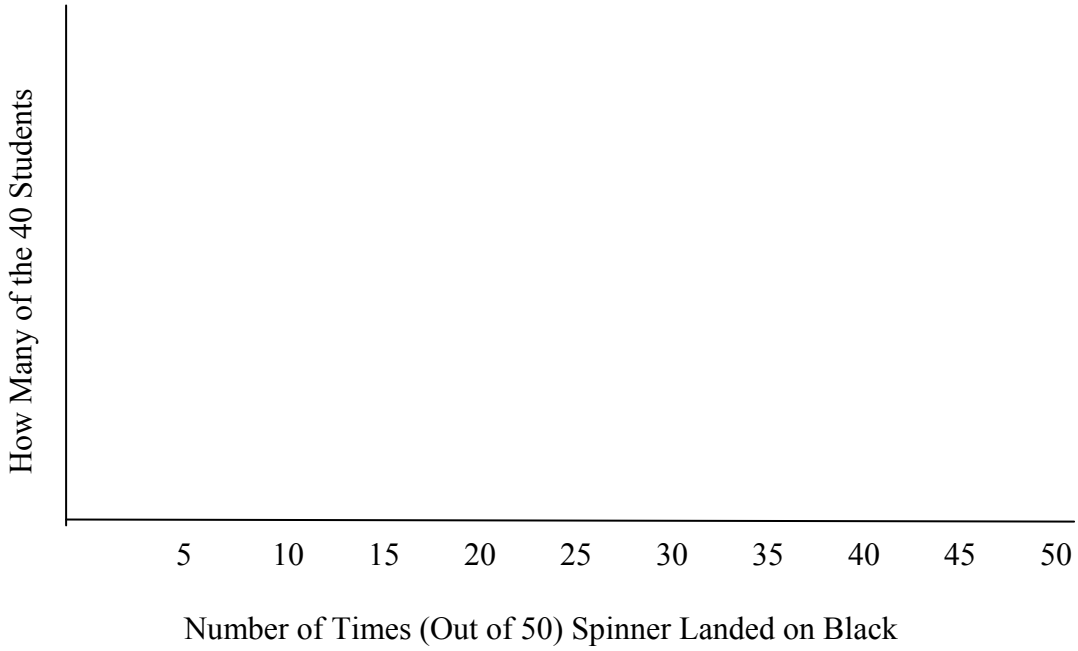
Now suppose 300 students did this experiment. What do you think the numbers (of times the arrow lands on black) will most likely go from?

From a low of _____ to a high of _____.

Why do you think this?

[3] Forty students lined up at the spinner. Each student spun it 50 times and wrote down how many times (out of 50 tries) the arrow landed on the shaded part.

At the end of class, they decided to make a graph of their data. Show below what their graph might look like:



POSTINTERVIEW

Second Interview Date: Name:

[1] Suppose there is a large container with 1000 pieces of candy in it. 600 are Red, 400 are Yellow. The candies are all mixed up in the container.

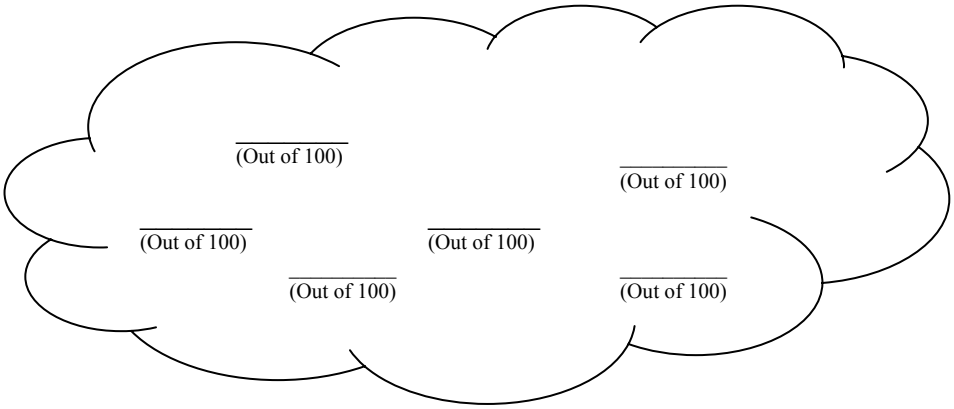
You reach in and pull out a handful of 100 candies at random.

(a) How many red candies do you think you will get?

(b) Suppose you do this several times (each time returning the previous handful of 100 candies and remixing the container). Do you think this many reds would come out every time?

Why do you think this?

(c) Suppose six classmates do this experiment (each time returning the previous handful of 100 candies and remixing the container). Write down the number of reds that you think each classmate obtained:



Why did you choose those numbers?

[2] Here are some examples of what other people have said for the numbers of reds that they think the six classmates would get in each handful. Put a check mark next to any of these that you think might be likely:

i) 72, 91, 74, 63, 81, 78

ii) 61, 73, 56, 69, 59, 48

iii) 60, 60, 60, 60, 60, 60

iv) 53, 41, 34, 60, 46, 52

v) 61, 66, 62, 62, 60, 59

vi) 30, 10, 90, 20, 60, 50

(a) Put a check mark next to any of these that you think might be likely.

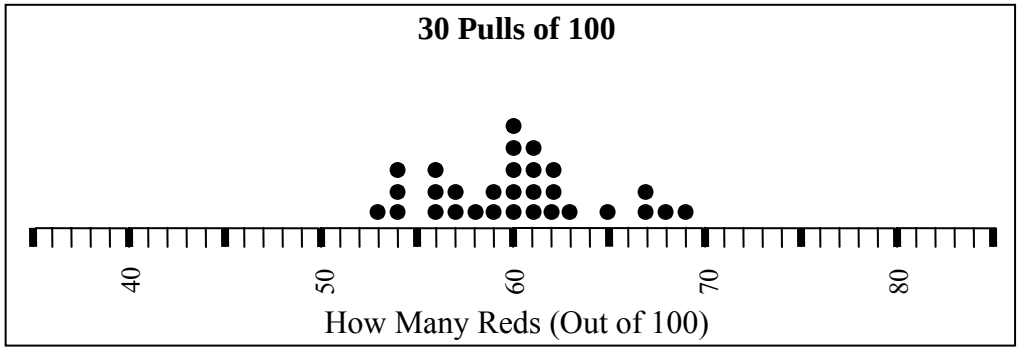
(b) Circle the list that you think best describes what might happen.

(c) Why do think the list you chose best describes what might happen?

[3] Craig took his class to a large candy container (1000 Candies = 600 Red and 400 Yellow). Then he left the room. When he came back, the class claimed to have pulled 30 samples each of size 100, with replacement. They showed Craig their data and a graph:

Number of Reds in 30 Samples of Size 100

53	54	54	54	56	56
56	57	57	58	59	59
60	60	60	60	60	61
61	61	61	62	62	62
63	65	67	67	68	69



Which of the following do you think is *most* likely ? Put a check mark next to it.

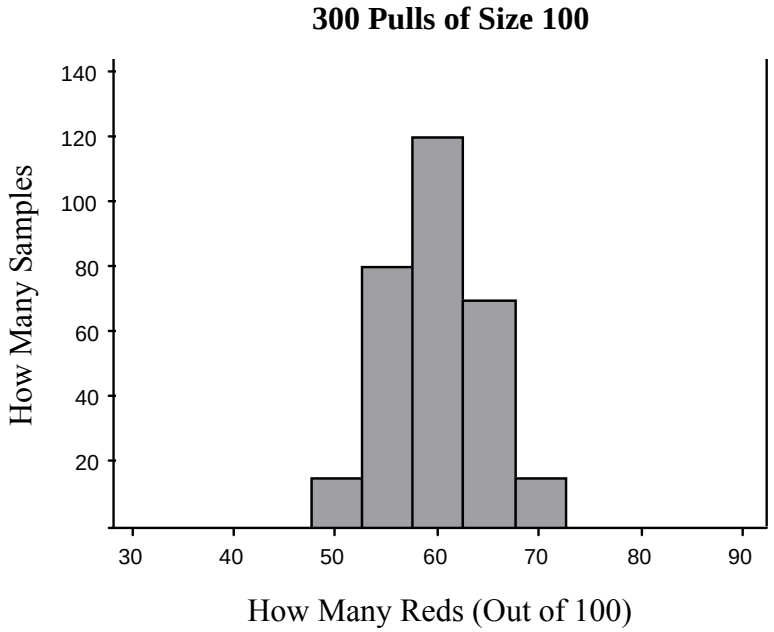
___ Craig's class just made up these results

___ Those are the actual results of the class samples

___ No one can have much confidence if the results are made up or not.

Explain why you think this is the most likely.

[4] Marj also took her class to a large container (1000 Candies = 600 Red and 400 Yellow). Then she left the room. When she came back, the class claimed to have pulled 300 samples each of size 100, with replacement. They showed Marj this graph:



(a) Which of the following do you think is *most* likely ? Put a check mark next to it.

___ Marj's class just made up these results

___ Those are the actual results of the class samples

___ No one can have much confidence if the results are made up or not.

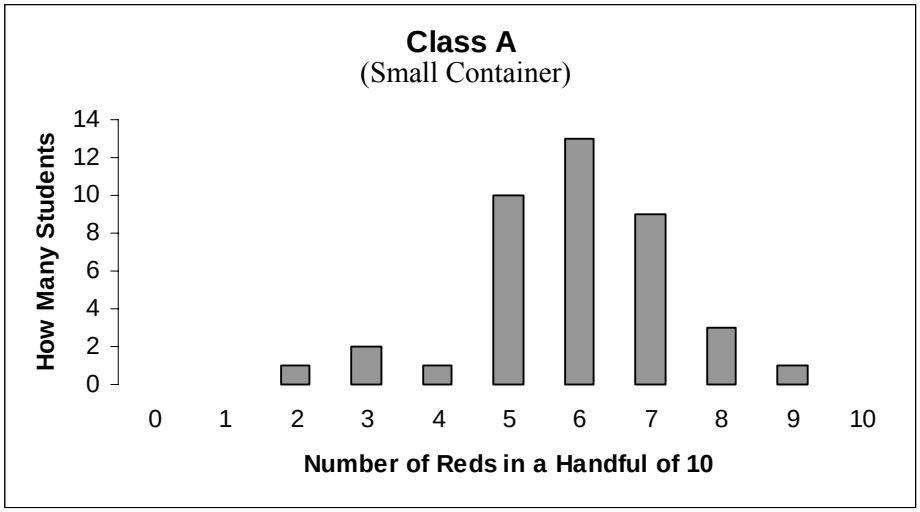
Explain why you think this is the most likely.

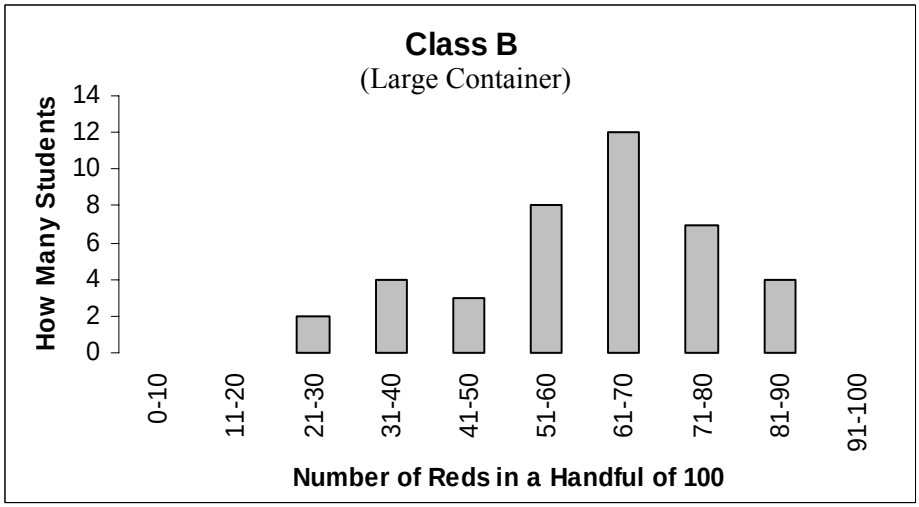
(b) How does the shape of the graph for Marj's class compare to the shape of the graph for Craig's class?

[5] On a day of planned absence from school, Keith left these instructions for his two classes:

- He told the forty students in "Class A" to go to the SMALL container (100 Candies = 60 Red and 40 Yellow). They were each supposed to draw small handfuls of 10 candies (with replacement after each draw).
- He told the forty students in "Class B" to go to the LARGE container (1000 Candies = 600 Red and 400 Yellow). They were each supposed to draw small handfuls of 100 candies (with replacement after each draw).

When Keith came back the next day, he saw these graphs and sets of data for the two classes:





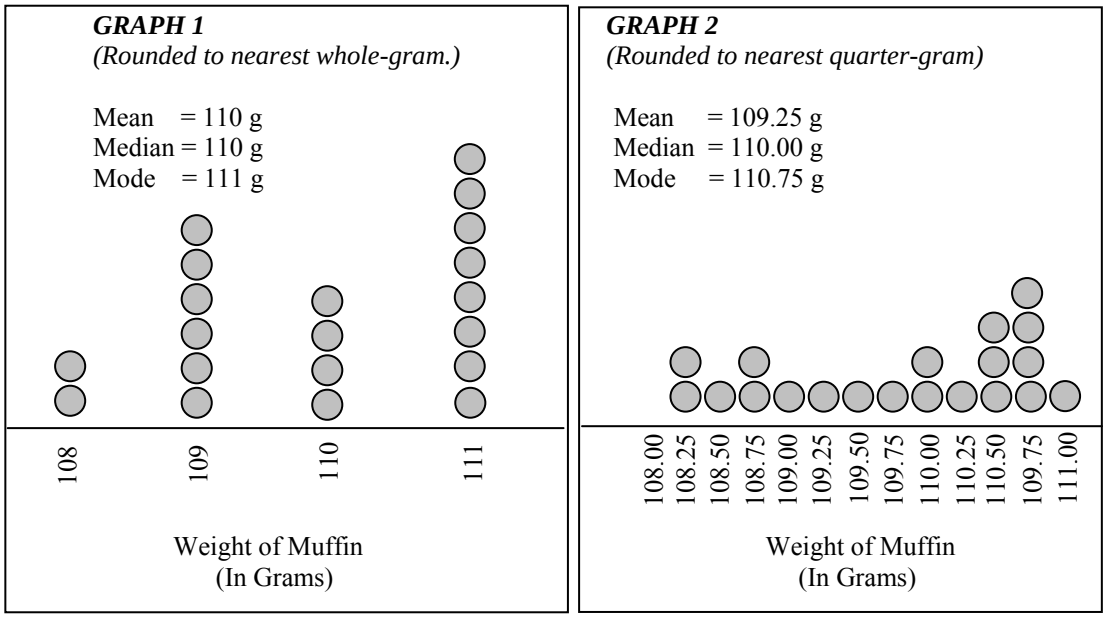
Keith suspects that one of the two classes just made up the data and didn't really carry out the experiment. What do you think? That is, based on the two graphs shown above, do think one graph is likelier than the other to reflect made-up data ?

[6] A class of twenty 6th-grade students wanted to find out the muffins weighed from some local bakeries. Their first goal was to see how good their measurement skills were, so they decided to find out the weight of a single muffin from the East End Bakery. They purchased one muffin, and took turns weighing it. They were surprised to find that they did not have the same results:

Weight of One Muffin (To Nearest 1/100 of a Gram)				
108.23	108.24	108.51	108.75	108.74
108.98	109.24	109.49	109.76	109.98
110.02	110.20	110.50	110.52	110.53
110.75	110.76	110.78	110.74	111.00

What are some possible reasons for why the class did not all get the same result?

[7] The class was deciding how to display their data. In Graph 1, they rounded to the nearest whole gram. In Graph 2, they rounded to the nearest ¼ - gram.



(a) How do these graphs differ in the stories they tell about the weight of the muffin ?

(b) Some members of the class argue that the muffin was really under 110 grams, while some argue that it was over 110 grams. Others claim it was exactly 110 grams. What do you think about the true weight of the muffin, and why do you think this?

(c) Does one graph help you more than the other in making your conclusion?

[8] A different class of 35 students wanted to know how much a "Grande Muffin" from the West End Bakery weighed.. They decided that each student should buy one "Grande Muffin" at different times during the week, and weigh the muffins before eating them.

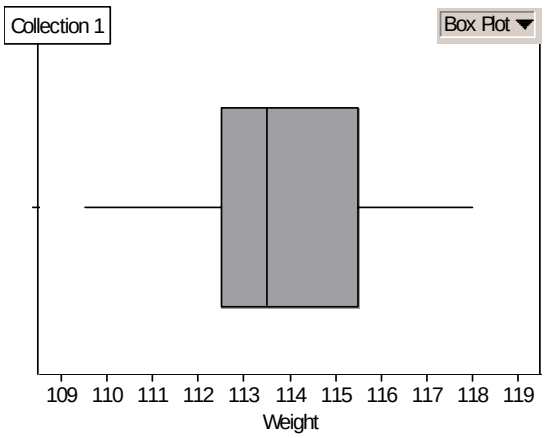
They recorded their 35 muffin weights (rounded to the nearest half-gram, see below), and summarized the results with a boxplot and a histogram:

Data: (Weight of Muffin)

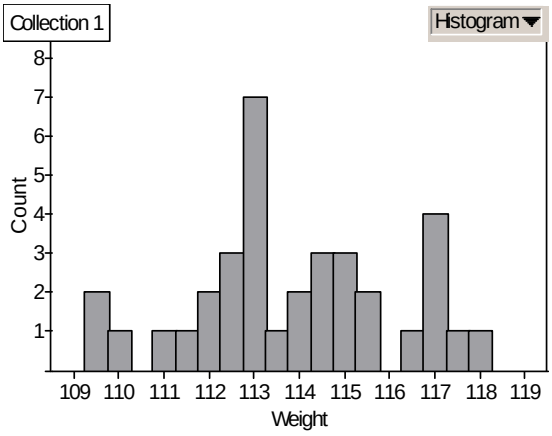
109.5	109.5	110.0	111.0	111.5
112.0	112.0	112.5	112.5	112.5
113.0	113.0	113.0	113.0	113.0
113.0	113.0	113.5	114.0	114.0
114.5	114.5	114.5	114.5	115.0
115.0	115.0	115.5	115.5	116.5
117.0	117.0	117.0	117.5	118.0

Mean	= 113.79 g
Mode	= 113.00 g
Median	= 113.50 g
Minimum	= 109.50 g
Maximum	= 118.00 g

Boxplot: (35 Muffins)



Histogram: (35 Muffins)



a) Suppose you bought a "Grande Muffin" from the West End Bakery. How much do you think your muffin would weigh? Explain your reasoning.

b) What are some similarities and differences in the way these two graphs present the data?

c) Do you think one graph tells you more about the variation in the data than the the other graph? Explain your thinking.

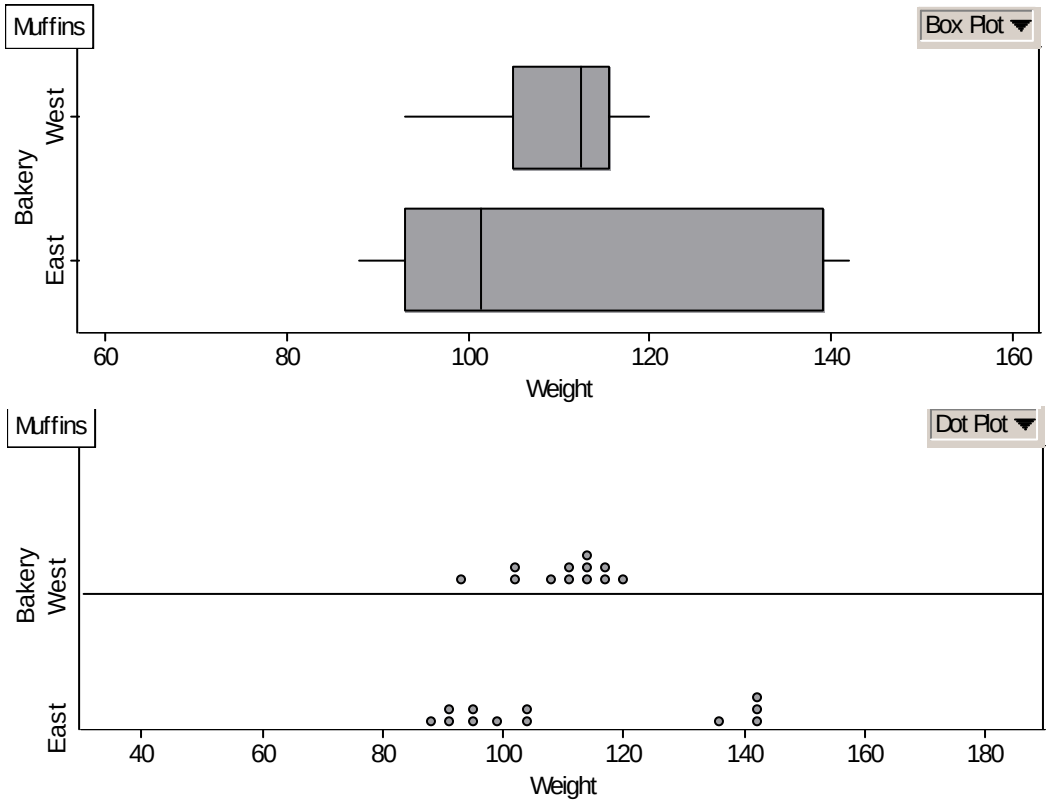
[9] The MathTeam must choose a new bakery to supply them with muffins. The MathTeam wants muffins that are usually at least 110 grams in weight.

So, the MathTeam samples twelve muffins from the West End Bakery and twelve muffins from the East Side Bakery. The weight of the muffins from the different bakeries were as follows:

West:	93	102	102	108	Mean =	110.25
	111	111	114	114	Median=	112.50
	114	117	117	120	Mode =	114.00

East :	88	91	91	95	Mean =	110.75
	95	99	104	104	Median=	101.50
	136	142	142	142	Mode =	142.00

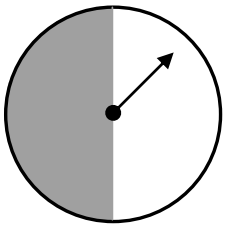
A Boxplot and Dotplot were used to portray the data:



i) Do you think that one bakery would better meet the needs of the MathTeam over the other bakery? Explain your thinking.

ii) Is one graph or the other more helpful in showing the differences between the two bakeries? Explain your thinking.

[10] Consider the spinner on the right:

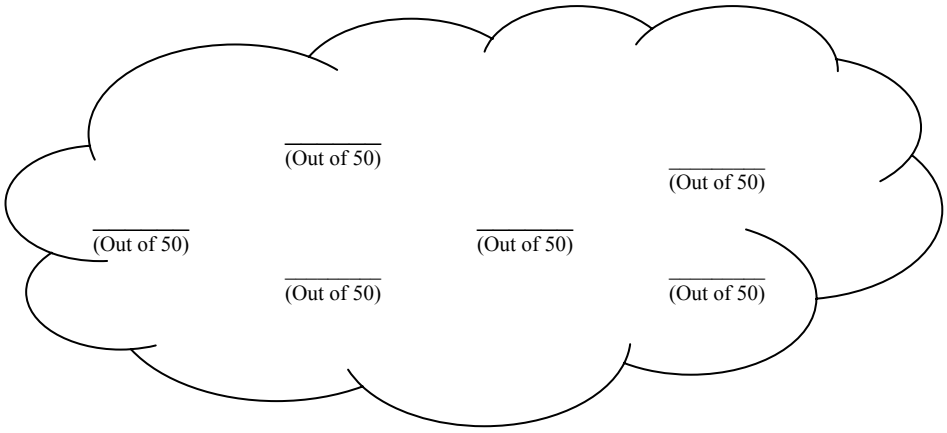


a) Matt is curious to see how often the spinner lands on black, so he spins it 50 times. How many times (out of 50 tries) do you think the arrow might land black?

Why do you think this?

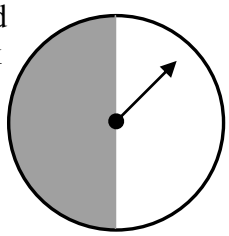
b) After Matt's first set of 50 spins, he decides to do a second set of 50 spins. How do you think his results on the second set of 50 spins will compare with the results of his first set?

c) Matt actually has a lot of time on his hands, so the next day he does 6 sets of 50 spins. Write a list that would describe what you think might happen for the number of spins out of 50 the spinner would land on the shaded part in each of the 6 sets of 50 spins.



Why did you choose those numbers?

[11] Here are some examples of what other people have said for the number of times (out of 50 spins) the arrows would land on black in each of the 6 sets of 50 spins. Put a check mark next to any of these that you think might be likely:



i) 38, 43, 36, 26, 41, 33

ii) 26, 32, 22, 29, 24, 19

iii) 25, 25, 25, 25, 25, 25

iv) 15, 19, 11, 25, 21, 23

v) 24, 25, 26, 25, 24, 26

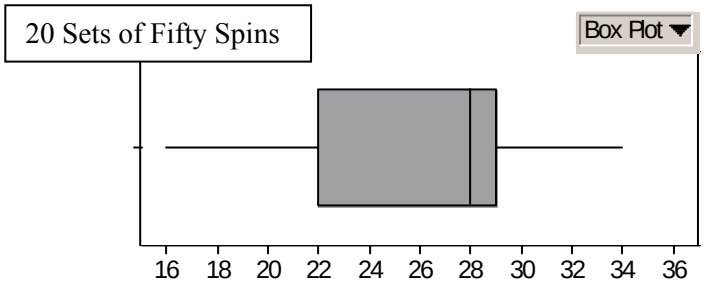
vi) 30, 10, 45, 20, 25, 35

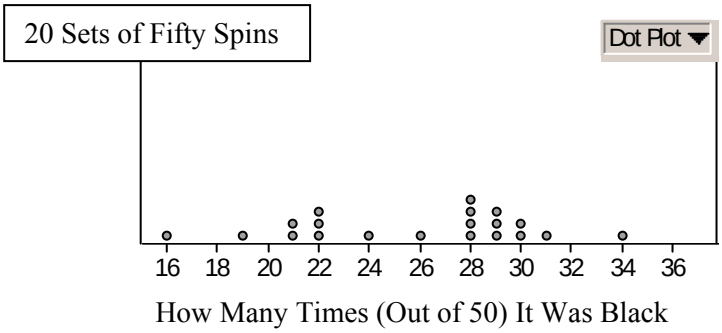
(a) Put a check mark next to any of these that you think might be likely.

(b) Circle the list that you think best describes what might happen.

(c) Why do think the list you chose best describes what might happen?

[12] Twenty students lined up at the spinner. Each student spun it 50 times and wrote down how many times (out of 50 tries) the arrow landed on the shaded part. At the end of class, they decided to make graphs of their data. Their boxplot and dot plot looked like this:





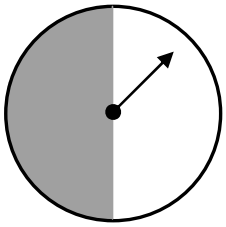
* Keith argued that something was wrong with the experiment because no one got exactly 25 out of 50 spins landing on black.
* Karen argued that, since the mode (and median) was 28 out of 50 spins landing on black, something was wrong with the spinner.
* Jeanette argued that the maximum of 34 out of 50 spins landing on black would not have happened unless something was wrong with the spinner.
* Marjorie argued there was nothing wrong with the spinner, since she had expected the results to look like this.

(a) What do you think about each person's argument?

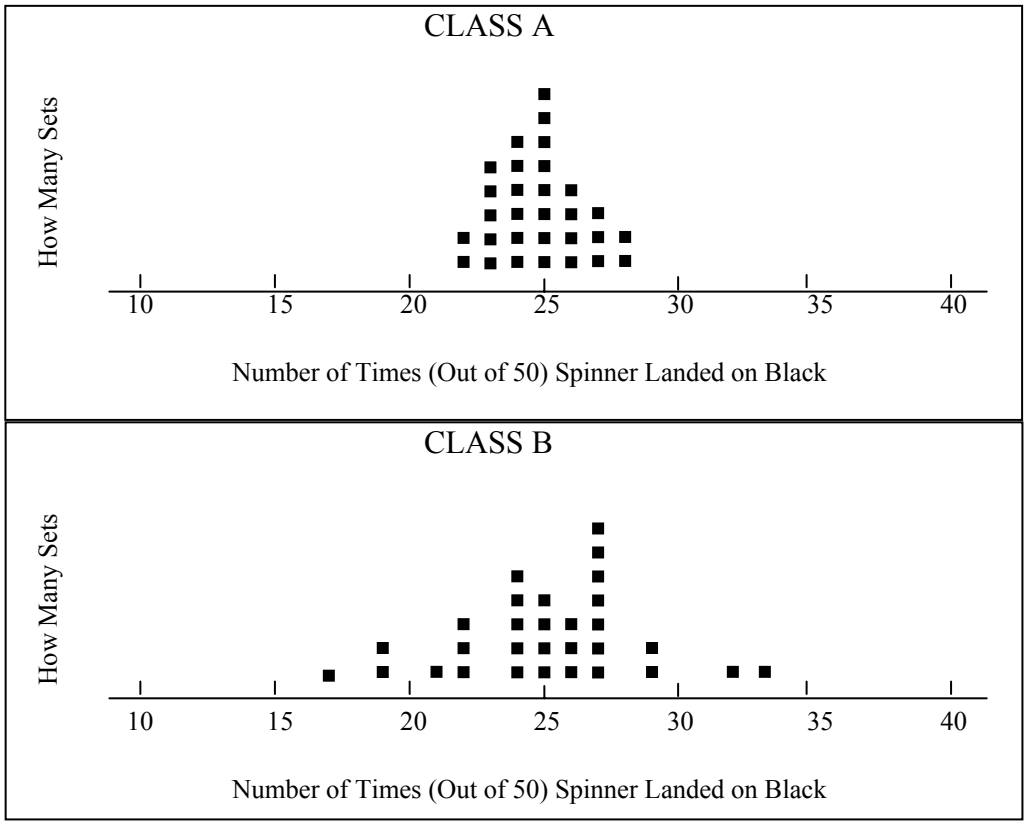
(b) Is there a different argument that you would make?

Explain.

[13] On a day of sick leave, Mr. Shaw left instructions for Class A to conduct 30 sets of 50 spins. For each set of 50 spins, they were supposed to record how many times (out of 50) the spinner landed on the shaded part. Then they had to graph the results of the 30 sets. He left these same instructions for another of his classes, Class B.



When he came back the next day, he saw these two graphs, showing the results of 30 sets of 50 spins for Class A and Class B:



(a) How would you describe the shape of the graph for Class A?

(b) How would you describe the shape of the graph for Class B?

(c) Mr. Shaw suspects that one of the two classes did not really do 30 sets of 50 spins, but instead just made up the data. Based on the results shown in the two graphs, do you think one class or the other is likelier to have simply made up the data?

Explain why you think this.

APPENDIX C

Class Interventions

Class Intervention #1: Data & Graphs

In the fifth week of the quarter, Steve left the topic of geometry and entered the topics of statistics and probability. Matt and I also began attending class, and we set up our audiovisual equipment on a table in the back of the room, leaving the class with seven table groups' worth of students instead of the previous maximum of eight. Steve's section had an enrollment of 30, so after Matt and I began attending the groups at the remaining seven tables had four or five students in each group. I was in attendance for eight days (spanning weeks 5 through 8), and some of the main activities

The contexts of data and graphs, sampling, and probability shared considerable overlap, and in fact Steve began with a class discussion of each of the terms which he had written on the board: Data, statistics, and probability. I have identified for each intervention a context which I thought had a more unique focus, and the dominant context in which Steve first led the class was data and graphs. The two activities comprising the Class Intervention for the context of data and graphs were called "Four Questions" and "Body Measurements," and are discussed next. It is important to remember that the purpose in sharing the details of these class interventions is to create a picture of the opportunities given to the subjects for exploring variation. Shifts in thinking from the pre- to post-instruments may indeed be attributable to the interventions, but this research does not set out to prove a treatment-and-effect dynamic. However, the environment for learning is certainly important to document, as it does offer clues as to how conceptions may be formed and influenced.

The "Four Questions" activity, as a part of the first intervention, was chosen because for two reasons. One reason is because Steve and I had each used versions of the activity with other Math 212 classes, and were therefore experienced in how it went and what it offered. The second reason is because it offered a good opportunity to discuss both average and spread in data sets. Steve therefore started the class exploration of statistics in the fifth week by having the entire class gather data from one another in response to four questions:

How many pets do you have?
How many years have you lived in Portland (to nearest half-year) ?
How many people are in your household?
How much change (in coins) do you have today?

Different groups were in charge of graphing the collected data for an assigned question in any way they wanted. Because there were more groups than questions, some questions were duplicated by different groups. However, being a very open-ended activity, the same question ended up being graphed by different groups in different ways. The graphs were all put up on poster paper in front of the room, and comparisons were made between the different types of graphs. An example of an actual graph for the number of pets of shown below in Figure C1:

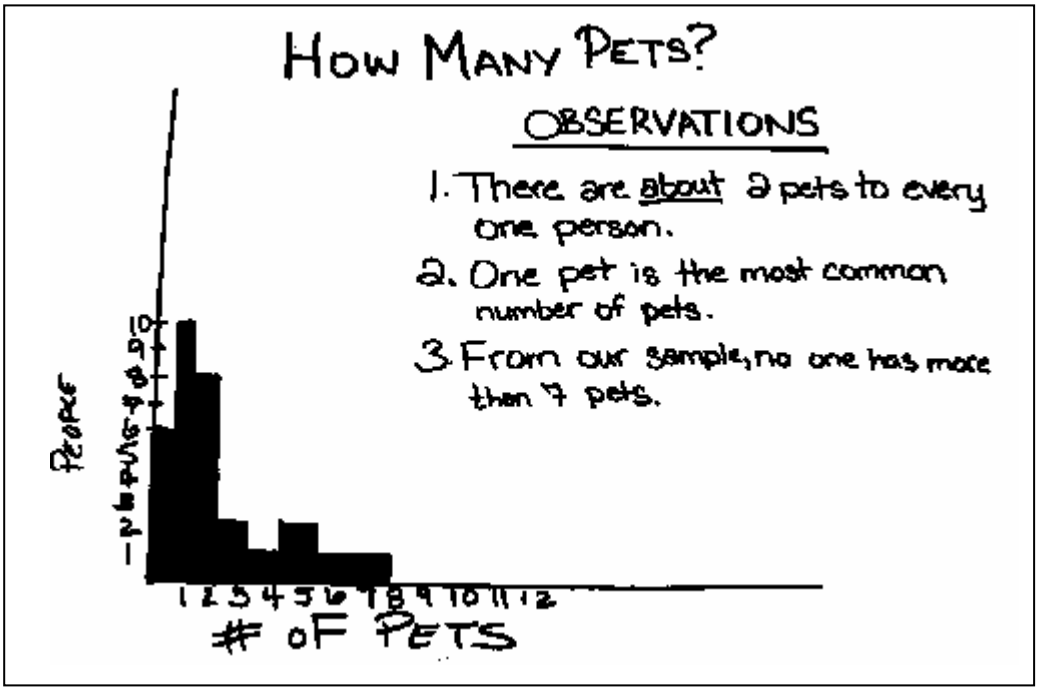


Figure C1 – How Many Pets?

As an example of how, for example, the contexts of sampling and data and graphs could overlap, here is an example of a comment volunteered by one of the six cases, SP, speaking on behalf of her group about a graph showing the number of people in a household:

> SP: The most common [number] was two…No household contained more than five, and that surprised us. So we thought, "Why is that?" And then we were thinking that this class is probably not representative of the population as a whole, because of age, education level, academic status…

The idea of representativeness clearly came through from other groups also.

Having done the activity with different Math 212 classes, the variety of types of graphs the students came up with on their own was greater than I'd seen with other classes. Steve's section had come up with pie and bar charts, histograms, pictographs,

and line and dot plots. A discussion ensued about the different ways that data could be presented, and also about the level of detail provided by each type of graph. One line of questioning that Steve and I prompted was the idea of what would be a "typical" value for a Math 212 student or for the class. For instance, the number of years living in Portland ranged from less than one year to over 40 years (see Figure 2). We could find an average, but in what sense would that value be typical?

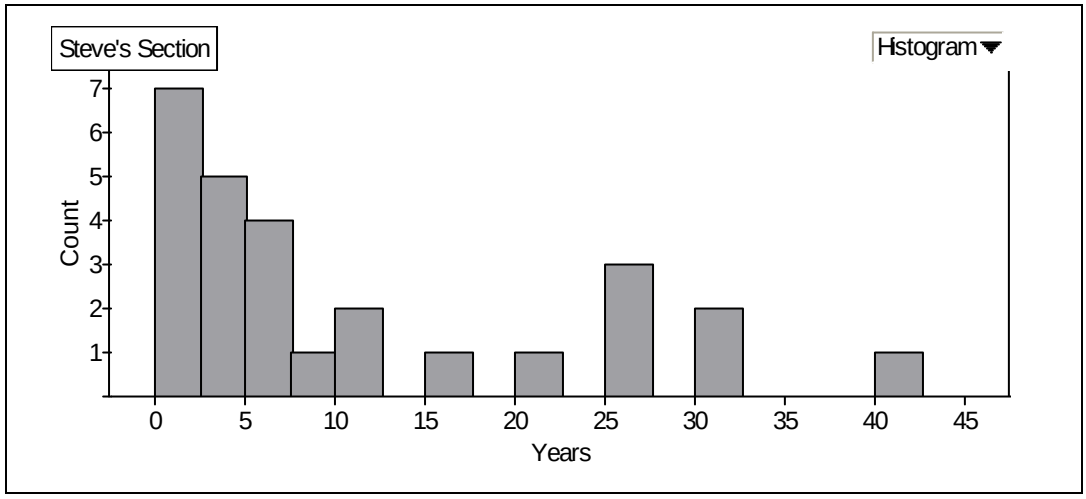


Figure C2 – Years Living in Portland

Figure C2 represents a computer-generated version of one kind of graph the class had in the front of the room for discussion purposes. Another table used a pie chart for displaying the data for number of years in Portland. The ideas that Steve was able to draw out or introduce included the definitions of mean, median, and mode, and how those measures appealed to a sense of average. He also highlighted the term "outlier" for the maximal value. The class picked up on how a measure like the mean doesn't even have to be an actual data point, as in Figure C2, where the mean is 11 years. The class expressed some dissatisfaction in claiming, for instance, that the typical student from Steve's section had lived in Portland for 11 years, when (a) No one actually reported the value of 11 years, and (b) More responses seemed to be under 11 years as opposed to over 11 years (the median in Figure 2 is in fact 5.5 years).

The main point is that the tension between centers and spread of data emerged as theme of discussion even from the very first day of doing statistics. Steve then went on in the next class session to investigate a model for obtaining the mean by balancing out stacks of tiles, and he also worked with the class to create and use boxplots. Particular attention was paid in the class discussions to what kind of information boxplots showed and what kind of detail they obscured.

The second activity in the class intervention focusing on the context of data and graphs was "Body Measurements", which was also selected for the reason that it was a well-rehearsed activity for Steve and me. More importantly, a similar activity

was to be used for the NSF-sponsored project (mentioned earlier) that looked at conceptions of variation in middle and high school classes. Having been a part of that project, Matt and I had agreed that the activity was useful for prompting thinking about variation. In particular, causes of variation in a repeated-measurement scenario get explored in Body Measurements, as do issues of confidence in situations involving measurement.

The introduction to Body Measurements came towards the end of the first class session of week 6, with about 30 minutes left in class. By way of introduction, we discussed scenarios that we'd already seen in class whereby the data varied. For example, the data in response to the four questions shared earlier had varied. We also listed things we thought would not vary, such as the time class started each day, or the amount it costs to park on the street for an hour during classtime. I offered the idea that some of our body measurements – such as armspan or head circumference – would not vary during a short interval of time. Thus, the task was given to gather several sets of data: First, everyone in class was going to measure Matt's armspan. Second, everyone had a "personal data sheet" which required their own armspan, height, handspan, head circumference, and pulse rate per minute. The only directions given were that pulse had to be counted out for a full minute, not some shorter interval and then multiplied by an appropriate factor.

What happened next mirrored in many ways what Matt and I saw later in the middle and high school classes. Armed with meter sticks and tape measures, the Math 212 students did find partners to help gather their own personal data, and in many cases the measurements were carried out very casually. For example, head circumferences were measured above the ears for some people and around the ears for others. Handspans included a natural span (meaning the subject just opened a hand as far as it would naturally go) or a forced span (meaning that the subject found some way to open their hand even farther, such as pushing the open hand against a table). What was really interesting was the measuring of Matt's armspan. Since none of the measuring tapes or meter sticks would singularly cover the armspan, subjects were forced to find a way of compensating.

The typical way of measuring was to start at one side and measure across Matt's back until, for instance, the measuring tape ended. Then subjects would hold a finger at the ending spot, or even affix their gaze to the ending spot, and then start a new measurement from that spot to the other side. For the twenty-seven students that we had in attendance that day, the time needed to gather the measurements was sufficiently lengthy to make Matt's outstretched arms sag after awhile. Eventually he would just put up one arm at a time as needed. There were so many sources of error in just the 15 minutes of data gathering that it was not possible to list them all, but it is interesting that many of the same issues arose later when Matt and I watched the activity done in the middle and high schools. I collected all the personal data sheets, which also had each subject's measurement of Matt's armspan, and then class ended for the day.

Because the students had already practiced making different kinds of graphs in class, I decided to type up the class results and distribute sheets of the data and some graphs during the next class session. The main part of the class discussion was on interpreting the graphs. For example, the graph for Matt's Armspan is shown in Figure C3 below. Questions that we asked of the students included: Why are the measurements different? What can you conclude about Matt's armspan? How confident are you about Matt's true armspan? Several comments emerged to show that students knew many causes of variation in the repeated-measurements situation. As to conclusions about Matt's true armspan, there were different ideas expressed with different degrees of confidence. For example, the whole class seemed very confident that the true measurement was captured within the range, and some members felt that the mean (about 76.5 cm) was the true value. Other students liked the mode (which was also the median value of 76 cm) because they felt that most students must have done the measurement correctly.

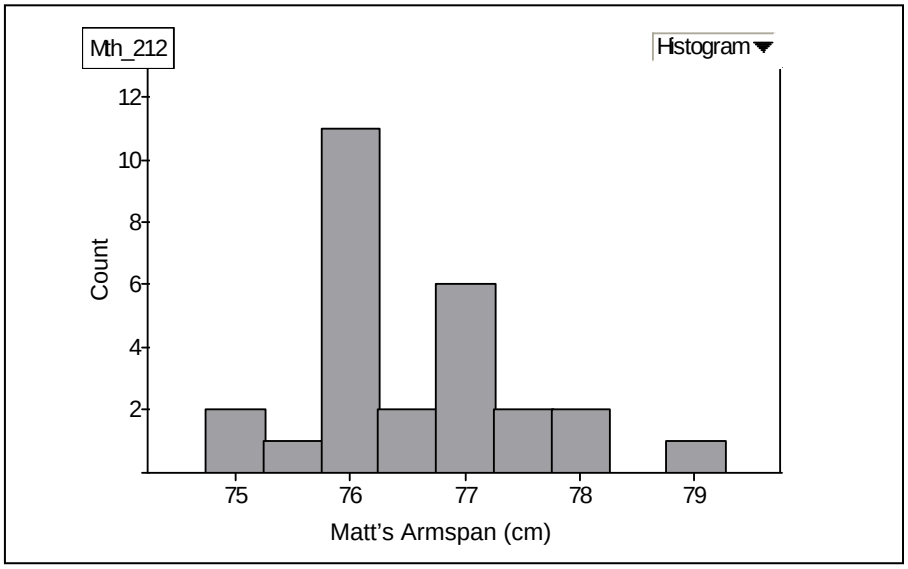


Figure C3 – Matt's Armspan

A big idea to come out of the discussion of Matt's armspan was that it was harder to identify a single value as being true or "accurate", but it was easier to talk about intervals of values where the true value might lie. Then in talking about the armspans for the whole class, again there was a similar discussion as we had for the "Four Questions" about what was a typical measurement for the class. Only a couple of students made a connection between the variation shown for a single measurement (Matt's armspan) and what that variation implied about the accuracy of the classwide measurements. The comment that came out in class were to the effect that, since Matt's armspan was "fuzzy" (meaning hard to pin down with precision based on the data), so too should each of the classwide measurements be "fuzzy" as well.

Class Intervention #2: Sampling

In the seventh week of class, we spent most of both class sessions doing two activities that comprised an intervention focused on the context of sampling. The two activities, "Known Mixture" and "Unknown Mixture," were selected because Matt and I had participated in the activities at six other schools as a part of the NSF-sponsored project mentioned previously. We had seen how effective the activities were in drawing attention to variation, and decided to follow the same essential plan for carrying out the activities as had been enacted in the middle and high schools. The sampling activities had evolved in the NSF project through meetings with classroom teachers and university researchers with an interest in promoting thinking about variation. Steve had done similar activities in previous Math 212 classes.

For the Known Mixture, we started with a general discussion of what were samples, who uses samples, and what samples were good for. Then the following scenario was given as a part of a handout (see Figure C4):

Scenario for Known Mixture Activity

The band at Johnson Middle School has 100 members, 70 females and 30 males. To plan this year's field trip, the band wants to put together a committee of 10 band members. To be fair, they decide to choose the committee members by putting the names of all the band members in a hat and then they randomly draw out 10 names

Figure C4 – Known Mixture Activity

Individually, the Math 212 students considered how many females they might get in a single draw of ten names (referred to a one sample). Then they wrote down what they thought they might get in six samples, and finally they made predictions for thirty samples. It was made clear that multiple samples were done with replacement. After making the above predictions individually, the students talked within their groups, coming to a group consensus about what they expected for 30 samples, and then they brought their numbers up to the overhead. I wrote all seven groups' numbers down, and they are shown in Table C1 on the next page.

In discussing similarities and differences in what the groups had predicted, students noted that all groups had highest numbers of females between 6 and 8, and the mode for almost all groups was at 7 (which corresponded to the proportion 7/10 = 70/100). Also, all groups had at least some results at 10. Three groups had no results lower than 3 females in a sample of size 10, while three groups predicted at least one sample having 1 female. After discussing the predictions, the directions for actually drawing 30 samples of size 10 were given.

Table C1. *Predicted Results for 30 Samples of Size 10*

Number of Females	0 F	1 F	2 F	3 F	4 F	5 F	6 F	7 F	8 F	9 F	10 F
Group #1	0	0	0	1	1	1	5	10	5	5	2
Group #2	0	0	0	1	2	3	6	8	6	3	1
Group #3	0	0	1	1	3	2	5	7	6	3	2
Group #4	0	1	1	2	3	4	5	6	5	2	1
Group #5	0	1	1	2	3	3	5	6	5	3	1
Group #6	0	0	0	1	2	4	6	6	5	3	3
Group #7	0	1	1	1	2	2	5	10	5	2	1

Plastic jars with 100 chips (30 green and 70 yellow) had been prepared, with the yellow chips corresponding to the females in the band committee scenario. Students were instructed, in their groups, to draw 30 handfuls (each of size 10) , recording the numbers of yellows in each handful before returning the chips to the jar for remixing. The subsequent sampling showed many of the same lackadaisical mixing techniques as had been observed in the middle and high schools. For example, some students would return their chips to the jar and then just give a weak side-to-side shake. Especially because the chips were flat and smooth, a sideways motion is not an optimum strategy for mixing. Other students would use the stirring technique, putting their hand in for a brief stir before drawing out their new sample. Another method was the up-and-down shake of the jar, which generally had to be tempered by the fact that the jars had no tops. Too vigorous of a shaking sometimes resulted in chips getting ejected from the jars.

Eventually, twin posters went up for each group: The top posted held the graph of that groups' prediction, and the bottom poster held the actual results. We were able to get all seven pairs of posters up on the boards in the front of the room. Figure C5 shows two of the posters.

In discussing the posters, the initial questions were "What do you notice?" and "What was surprising?". Some of the comments that followed had to do with the centers, spreads, and shapes of the graphs. For example, in comparing from predicted to actual, while almost all of the modes for the predicted graphs were at 7, only three of the actual graphs had modes at 7 (modes for the groups' actual graphs were 6, 6, 7, 7, 7, 8, 8). All of the actual data is shown below in Table C2.

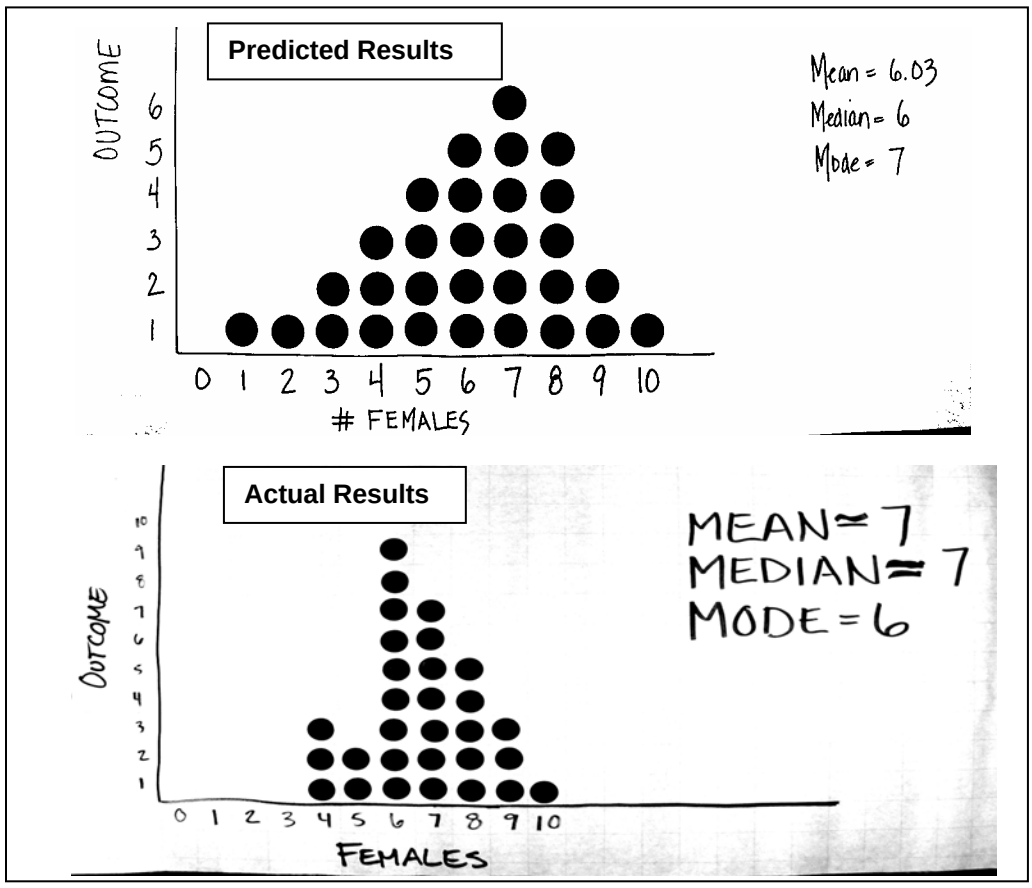


Figure C5 – Posters for Known Mixture

Students brought up how the predicted graphs were more spread out than the actual graphs, and also how the shapes of the predicted graphs all looked somewhat similar, but the actual graphs had shapes that were noticeably different from one another.

Table C2. *Actual Results for 30 Samples of Size 10*

Number of Yellows	0 Y	1 Y	2 Y	3 Y	4 Y	5 Y	6 Y	7 Y	8 Y	9 Y	10 Y
Group #1	0	0	0	0	1	3	8	5	6	7	0
Group #2	0	0	0	1	3	3	6	6	7	3	1
Group #3	0	0	0	0	2	4	4	7	6	6	2
Group #4	0	0	0	0	3	2	9	7	5	3	1
Group #5	0	0	0	0	0	4	7	9	6	2	2
Group #6	0	0	0	1	1	5	8	10	3	1	1
Group #7	2	0	0	0	0	5	4	6	8	4	1

On the issue of extreme values, it was mentioned how not all groups got a 10, and how most groups did not get below a 4. Group #7, who had reported getting two handfuls containing zero yellows, did raise suspicions among other class members. They asked the members of Group #7 if they had adhered to the procedures for drawing handfuls, , and it turned out that one of my cases, GP, admitted to using his sense of touch to discern the colors of the pieces. One of GP's group members, MG, said: "If you really compared the feeling of the two [types of chips], the yellow had a sharp edge and the green had a blunt edge." GP was also shown in the interviews to have an acute awareness of the role of his hand in creating (or impeding) randomness in these kinds of sampling situations. No one else in class seemed to have been aware of the tactile differences in the chips that GP's group had exploited to deliberately draw two handfuls of zero yellow in a row. Another of my cases, in a different group, asked GP if those two non-random samples had been calculated into the mean, showing an appreciation for the effects of outliers.

One big idea that came from the discussion was that the actual posters were not all the same, and in fact looked even more different from each other than had the predicted posters. Another big idea was that just as the predicted posters had most of the data around 6, 7, and 8, so too did the actual posters. Finally, disregarding GP's results of 0, a final big idea was that the lower extremes did not seem likely to result from actual results of 30 samples. At the end of the class session, it was suggested that we could get a better "actual" graph if we combined the results from our groups.

With these ideas in mind, the next class session began with an extended time of interacting with the ProbSim software, which some students later referred to as Fathom because that was the program I had used at an earlier time in the quarter. Matt led the class through much of the investigation as he had done with the NSF project. However, we spent more time using ProbSim with the Math 212 class than we had usually done in the middle and high schools.

The first thing we did was to run many trials of 30 samples. Since the students had done and seen the results for seven trials of 30 samples, the results (which ProbSim displays quite rapidly) made sense to the class: They were essentially imagining that hundreds of groups had done trials of 30 samples, not just seven groups as they themselves had done during the previous class session. Since we had mentioned aggregating our seven groups' worth of data, Matt shifted to trials of 210 samples (corresponding to seven groups at 30 samples per group). Two things that students were quick to pick up on was how the shapes started to look very similar from trial to trial (of 210 samples in each trial). In particular, the mode had stabilized at seven, and the graphs all looked like skewed bells. However, the far extreme of 1 yellow did not always result, and 0 never appeared. Matt then did repeated trials of 500 samples, and also 5000 samples, but 0 never appeared. Students felt sure that since 0 was a possibility, it should result given enough samples. Matt, Steve, and I pointed out that collectively we had seen tens of thousands of samples of size 10, and none of them had resulted in 0 yellow. There was a considerable anticipation in the

class, waiting for a result of 0, but we ended the computer simulation before that result occurred.

We then made a transition into the second activity in this intervention, which was the Unknown Mixture. It was made clear that even though we had known what was in the earlier jars, samples still had varied. Now we had larger jars, each containing 1000 chips of yellow and green, and the mixtures were known to be the same in each jar. However, the exact mixture was not known to the class (it was actually 550 yellows and 450 greens). The students were, in their groups, to decide what sample size they wanted to use (we imposed an upped limit of size twenty for all groups) and how many samples they wanted to draw. Then they were to carry out their plans, do the sampling, graph the results, and make some conjecture about the true mixture in the jar.

As an example of one groups' results and reasoning, the poster in Figure C6 shows how the mean and the median of 5 yellows helped this group decide on a prediction of 50% yellow in the jar.

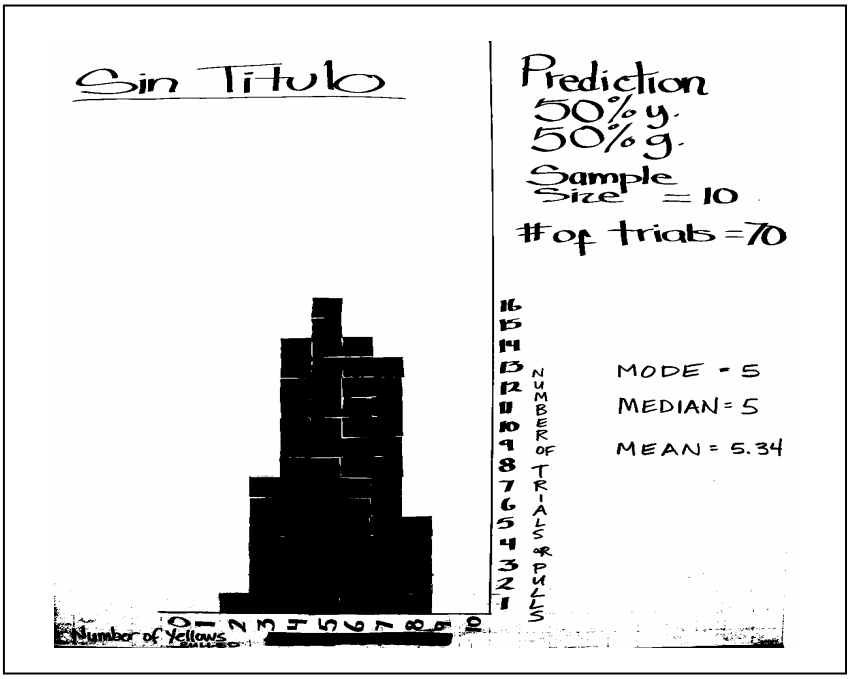


Figure C6 – Poster #1 for Unknown Mixture

The groups who authored the poster in Figure C6 used a sample size of 10, and they drew 70 samples. The group who made the poster shown in Figure C7 on the next page also used samples of size ten, and they drew 40 samples. However, they obtained their "guesstimate" of 57% by looking to their mean of 5.47 yellow and also their mode and median of 6 yellow, and finding some a value that they felt was somewhat close to both 5.47 and 6, namely 5.7. Then they used the ratio of sample size to population and determined that 570 yellow chips out of 1000 total chips would

correspond to 57%. They added the "margin of error" because they knew that plus-or-minus three percentage points would cover their mean, median, and mode.

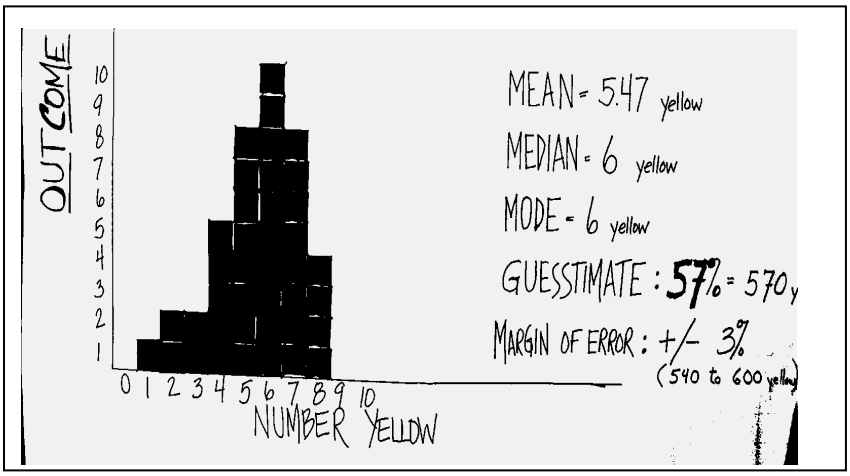


Figure C7 – Poster #2 for Unknown Mixture

Other groups selected sample sizes that went as low as 6 and as high as 200, and after discussing the different estimates we decided to try for a class consensus. I introduced the idea of confidence by relating the idea to what was shown in the posters. For example the group for the poster in Figure C6 had an estimate of 50% yellow, or 500 yellow chips, but they were not too confident the jars really did have exactly 500 yellows. The group for the poster in Figure C7 were not confident of the point estimate of exactly 570 yellow, but were fairly confident that the true value was somewhere in the interval from 540 to 600 yellow.

Comparisons across all the posters showed that the sample sizes for ranged from 6 to 20, and the numbers of trials ranged from 20 to 70. Predictions ranged from 500 to 600 yellows, with a couple of groups offering an interval. In discussion, I asked the class as a whole what would or would not be surprising to them: For example, the class expressed no surprise if the true value was 520 yellow, or 580 yellow. The class eventually came to a consensus on an interval as small as 540 yellow to 570 yellow. The big idea here was that an unknown mixture (or any other realistic sampling situation) does not mean that nothing can be said with any confidence about the mixture. In other words, the class was overwhelmingly confident that there were more yellows than greens, but not as high a ratio as 650 to 350, for example. Despite attempts to have the students accept that in real life sampling investigations, having some degree of confidence in an interval is the best that can be hoped for or expected, students still wanted to know the exact percentage, which was revealed at the end.

Class Intervention#3: Probability

There were two activities that made up this intervention, "Cereal Boxes"

and "The River Crossing Game," and these were chosen specifically because of the probability aspects involved in the activities. Namely, Cereal Boxes relies on the use of spinners and River Crossing on the use of dice as random generators, and these two activities were the main ones done in Math 212 involving spinners or dice.

Cereal Boxes actually took place in the first class session of week 2, just before we gathered data for Body Measurements. As explained earlier, there was considerable overlap in the contexts, and Cereal Boxes is a good example of this overlap. Cereal Boxes is sample-until scenario, supposing that any of five different stickers can be obtained with each box of cereal opened, and that the five stickers have equal chances of being obtained. The question is, about how many boxes would be opened to obtain all five stickers, and the situation can be simulated by using a five-region spinner. Cereal Boxes brings together probability, sampling, and data and graphs in way that also highlights variation, and that is why I chose to use the activity.

After posing the problem, guesses were taken from the class about how many spins they expected to have to do to hit every region at least once. An upper limit of 1000 was suggested, and one student suggested the upper limit was infinite. Also, a students suggested an expectation of 5 spins, the lower limit. Although this latter student was labeled an optimist by classmates, it focused attention on the probability aspects of this activity: After all, if we say there is a 1-in-5 chance of getting a region on the spinner, then after 5 spins there might be an expectation that all regions have been hit. As for the suggestion of infinite spins, it also focused attention on the idea that there was a chance of never hitting a certain region.

After discussing expectations, working individually at their tables, each student performed 10 trials (one trial was defined as finding the least number of spins it took to hit all five of the regions at least once). SW, for example, whose initial guess was 105 spins, obtained the following results for 10 trials: 9, 5, 20, 6, 14, 17, 6, 14, 10, 9. I asked her if, after looking at the results of her 10 trials, 105 spins was still about what she would expect:

> SW: I think I might lower it a little bit, maybe to 70, in the 70s. But I think it would take quite a bit, it might even be higher than that [105]. Because, I mean, you're not just dealing with numbers, then, you're dealing with, you know, they want you to buy as many boxes of cereal as possible, so they're going to spread it out a little bit more.

The class did raise the issue of what a cereal company's actual distribution would be like, and Steve re-focused the attention on the assumptions of the simulation. Thinking strictly in terms of numbers of spins to get the different regions on the five-spinner,

new expectations were made in light of the trials gathered. Instead of ranging from 5 to 1000 spins, the second round of predictions for an expectation ranged from 7 to 25.

Another key issue arose when Steve was asking the class to notice how the second round of predictions was tighter than the first round, and he asked "Which answer is right?" After someone said that there was no right answer, Steve steered the conversation to the idea that some predictions seemed more reasonable than others, and a part of the discussion included the idea of aggregating results to get even better predictions. After aggregating individual results at their tables, there were seven group predictions based on 40 to 50 trials, and the predictions ranged from 7 to 11 spins. Students did want to aggregate the whole class results sensing that more trials offered an increasing better idea of what to expect. Instead of taking the time to graph all the data, I brought out the Fathom software. Figure C8 shows the results for 150 trials:

Steve talked about where the upper and lower 10% of the data was, and also used the Fathom graphs to talk about boxplots and distribution of data. While the mode in the data set for Figure C8 is 9 and the median is 10, the mean is 11.7 and quite close to the expected value in this situation. But what students were aware of from having done their own trials was the variation in this situation, and how the chances of getting a certain region on that spinner were not guaranteed.

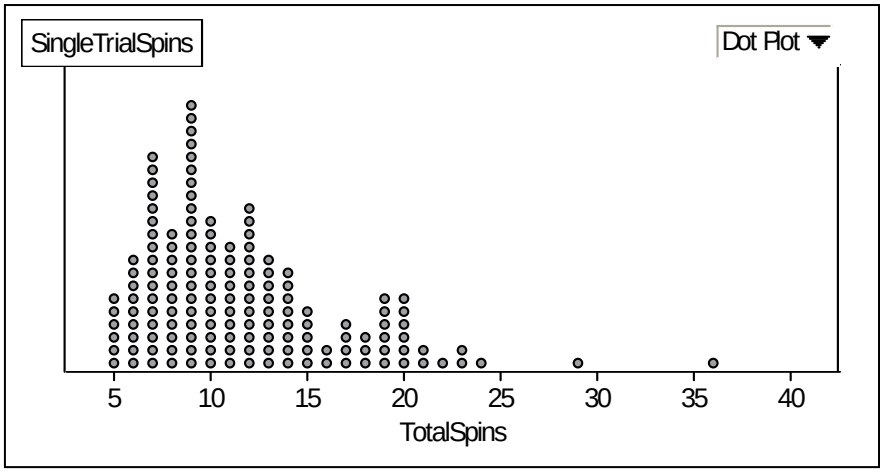


Figure C8 – 150 Computer-Generated Samples

In Figure C8, students knew that the meaning of the upper extreme was that one region on the spinner got avoided for 34 spins, only receiving a hit on the 35th spin. Cereal boxes, although modeling a sample-until scenario, did afford students good opportunities to focus on probabilities involving the spinners, which was among the reasons I had wanted to include it as a part of the intervention for the context of probability.

The second activity for this intervention, the River Crossing Game, involved finding the sum of two dice. Credit for this activity goes to the *Math and the Mind's Eye* curriculum (Shaughnessy & Arcidiacono, 1993). Using two players, each player got 12 chips to place on their side of a "river", along spaces marked 1 through 12.

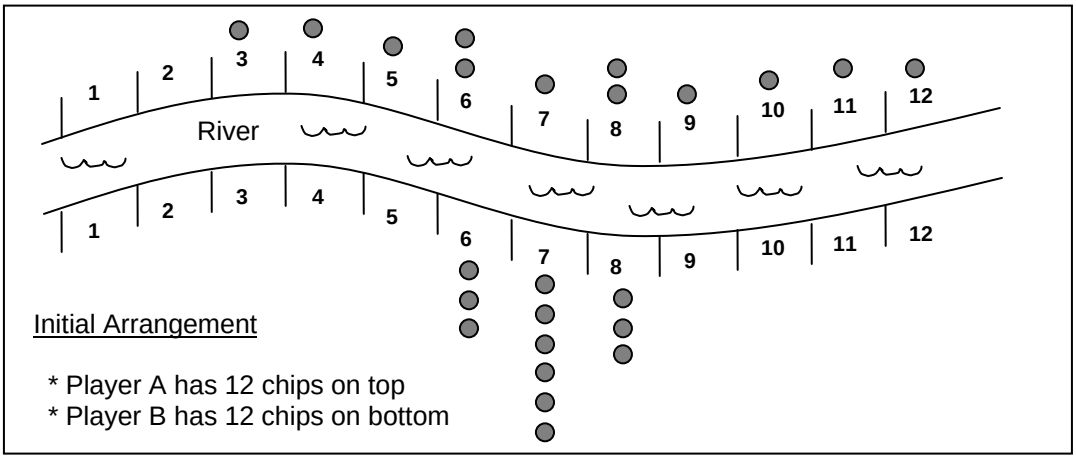


Figure C9 – Initial Strategies for River Crossing Game

After configuring their chips in an initial arrangement (see Figure C9 for an example of two players' initial arrangements), players took turns tossing a pair of dice. If either player had any chips on the space showing the total for the dice, one chip could "cross the river" and be removed from the board. The winning player was the first one to remove all the chips on his or her side. For instance, in Figure C9, if the dice resulted in a sum of 10, Player A on top could remove one chip. If the dice showed 8, Player A and B could each remove one chip.

Teams of students played several games, and kept track of the results of each toss of the dice on a dot plot: Sums were given along the horizontal axis, and students could just make a mark showing the sum obtained. Some students' initial arrangements changed from game to game. For example, seeing from their dotplots how sums of 6, 7 and 8 tended to occur more than other sums, some students put most of their chips on those spaces. Other students continued to spread out their chips, feeling that in the course of tossing the dice, they would get a sum of 2 or 12 for example.

After a few games, the dotplots from the teams were put up on the board in front of the class, and each graph showed well over one hundred tosses of the dice. Steve led the class in a discussion about the graphs, their shapes, and what the class as a whole might expect. Despite the variation shown in each graph, students volunteered that what they felt was most likely to happen for a sum would be 6, 7, or 8. Some students kept volunteering 7, but Steve made a distinction between the knowledge that some already held (that 7 was most likely) and what the class' experimental data suggested. Most of the class thought that 8 came through as being most likely, based

on class results. Steve pointed out that "there are trends, but there's still a lot of variation from group to group." Looking across the graphs, the idea of a symmetrical theoretic distribution also emerged, and the class later went on to look at the theoretical probability for the sum of two dice.

After the idea about 7 being the theoretically most likely sum emerged, Steve also discussed what the class thought might be an optimum initial arrangement for playing the River Crossing Game. An example of two strategies, informed by having now had experience playing and by looking at the theoretical distribution, are given in Figure C10 below:

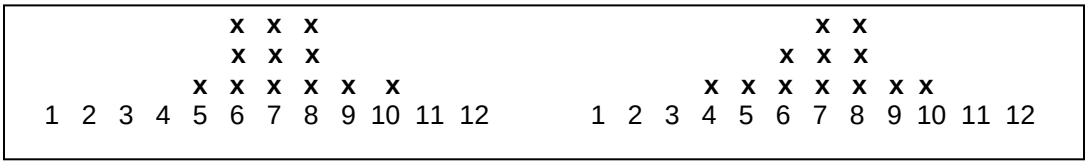


Figure C10 – Experienced Strategies for River Crossing Game

A question arose as to why people don't just put all their chips on 7, since that sum is most likely. Steve brought out the connection that putting all the chips of 7 was making a statement of expectation of no variation. He pointed out how the dotplots of results showed not only what happened most, but also gave an idea of the kind of variation we could expect as well.