CHAPTER SEVEN

Discussion and Conclusion

This chapter discusses how the study has addressed the research questions, and describes the main contributions of my research findings to the general field of studying probability and statistics education among teachers and students. The emergence of a detailed framework useful for looking at conceptions of variability is the key development to come out of the research analysis. The framework also informs the research questions by offering an in-depth exploration of a sample of elementary preservice teachers' thinking about variability.

Therefore, in the first section that follows, the emergent framework is discussed, bringing together the main results from analysis of both the survey and interview data. Then, in the second section of this chapter, the framework is used to look at a cross-case analysis of interview data, highlighting some similarities among and differences between the six cases interviewed. The third section summarizes results of the study in relation to the original research questions, and the fourth section discusses limitations of the research. Finally, the fifth section outlines some implications for future teaching and research.

Emergent Framework

This section discusses the development of the emergent framework for looking at conceptions of variability, and connects the framework to the research questions. The research questions for this study were largely

motivated by past studies, meaning that other researchers had focused on related questions but not with elementary preservice teachers (EPSTs). My research questions were:

1) What conceptions of variation are held by EPSTs in the three contexts of data sets, sampling, and chance situations?
2) How can the conceptions of variation held EPSTs in these three contexts be characterized?
   - A) What variation do EPSTs *expect*, prior to seeing the data or carrying out an experiment?
   - B) How do EPSTs produce and reason about *displays* of variation?
   - C) How do EPSTs *interpret* variation in terms of its causes and effects?

As explained earlier, the contexts of data sets, sampling, and probability (chance situations) were used because that is where previous research had focused. Also, the choice to frame my research questions in terms of the aspects of *expecting, displaying,* and *interpreting* variation was also motivated by models of statistical thinking used in past studies. In particular, the tentative, initial framework posed at the end of Chapter Three incorporated different dimensions as part of each aspect (see Figure 51):

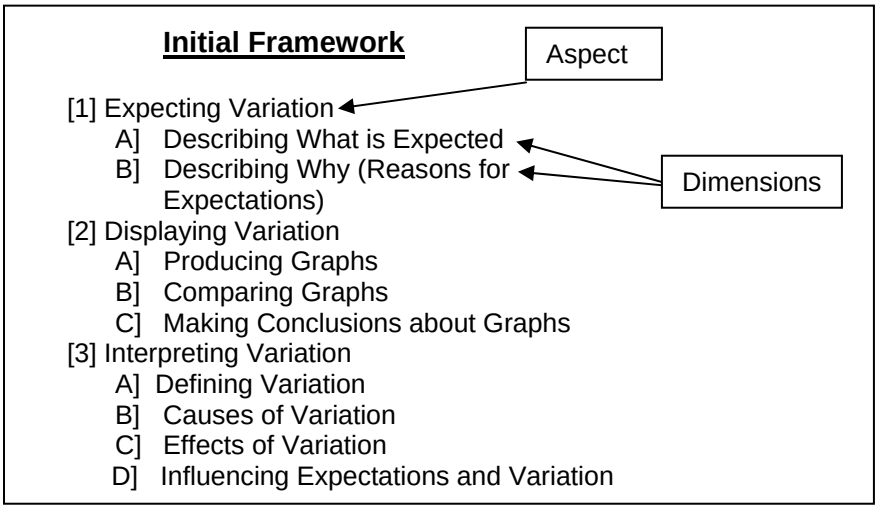

Figure 51

The working hypothesis going into the data gathering and analysis was that the initial framework could be useful for organizing and looking at responses from the students on tasks designed to prompt thinking about variability.

Taking the four surveys that were administered to the whole class, coding schemes were used to look at overall categories and major trends in responses. I also looked at the survey data to see how the responses might fit into and inform the initial framework. In addition to confirming the utility of the initial framework, the results from the survey analysis let me add detail, revising the framework so that I had a better sense of what themes were a part of each dimension. The emergent (revised) framework was carried into the analysis of the PreInterview data, where I considered responses from each of my six cases to each of the PreInterview questions. The analysis of the PreInterview data again added depth to the themes in the emergent framework, so that I was then able to consider the individual thinking of the six cases on a common subset of PostInterview questions.

The development of the emergent framework is therefore best seen in a recursive sense: As a lens for looking at conceptions, the initial framework helped me to see a large amount of classwide survey data in terms of some major aspects and dimensions. At the same time, what I saw in the classwide responses helped me sharpen the focus of the lens offered by the initial framework, so that some themes within each dimension emerged. The PreInterview data confirmed the utility of the themes and also added richness to the description of those themes, as did the PostInterview data.

Finally, after analyzing student responses while at the same time being informed by those responses, the emergent (revised) framework consisting of aspects, dimensions, and themes is presented in Figure 52 below:

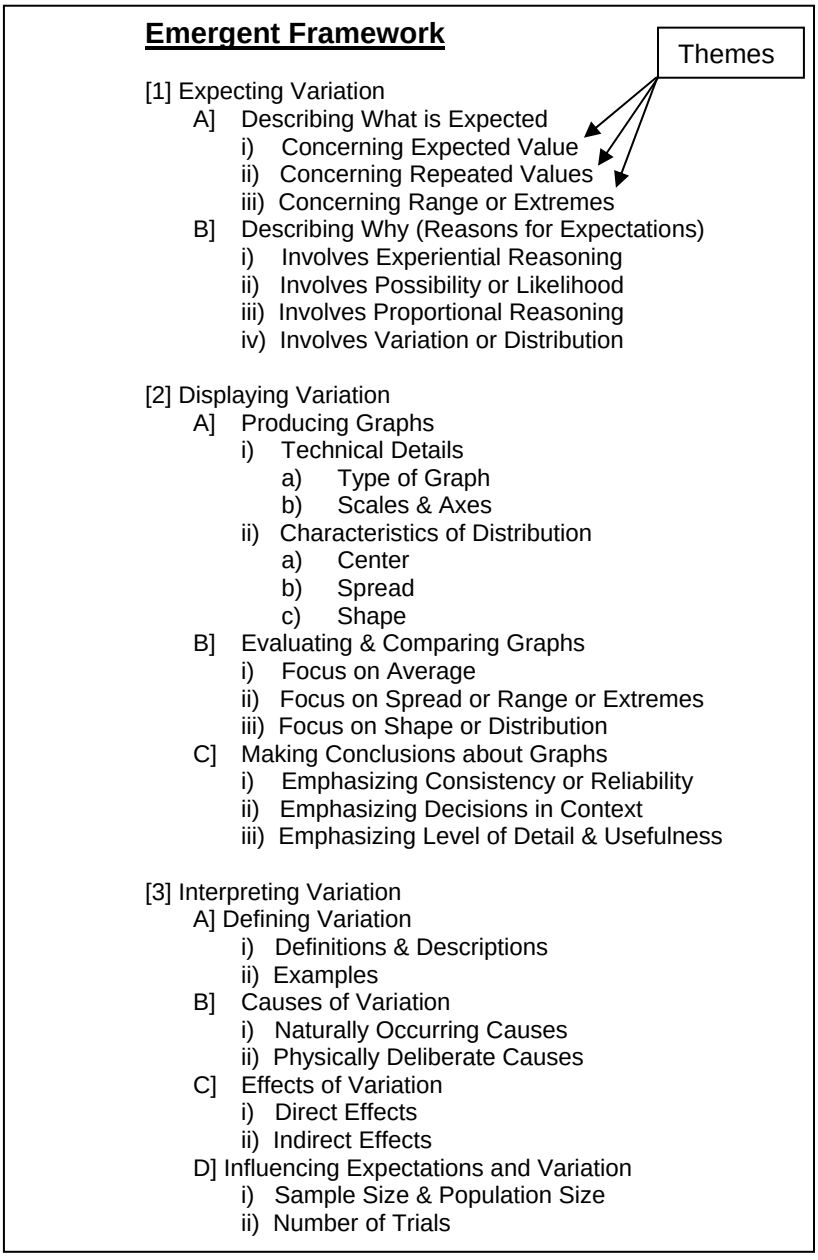

Figure 52

The emergent framework is hypothesized as a useful way to summarize the

conceptions of variation held by elementary preservice teachers in the contexts of sampling, data and graphs, and probability situations. The hypothesis is grounded in the data taken from both survey and interviews of EPSTs, and represents the key contribution of this research to the literature. As such, I'll summarize the descriptions of the themes within the framework, amalgamating the collective responses from survey and interview data to illustrate conceptions of EPSTs about variability.

[1] Expecting Variation

[1A] Describing What is Expected: The first theme [1Ai] concerning the expected value is characterized by responses which could emphasize that value in different ways. A dominant type of response in this theme was how results should be close to, about, or near the expected value. Subjects talked about expecting results that are both higher and lower than average, or how results shouldn't be the average each time, or even how results should cluster around the average. As has been mentioned, a student's response could span more than one aspect, dimension, or theme, due to the length and depth of a response. For example, DS had a response concerning the expected value in a sampling context on Q5 in the PostInterview ("Small & Large"), saying that "as you get farther from average number of reds – the 60 – then there are going to be fewer, fewer people drawing that number." She shows a sense of distribution as well, but her response also hinges on

the expected value of 60 reds in this situation. For responses concerning the expected value, some type of average seems critical to the overall reasoning behind the response.

Concerning repeated values [1Aii], some students suggested that that results would be likely to repeat more often than not, while other students emphasized how results would not be the same each time. When they made their own choices, such as on the “Six Trials” questions in the surveys and interviews, some students seemed to deliberately choose values so that all six choices were different from each other. An interesting example of a response concerning repeated values come from EM as she considered tossing a fair die on Q9 in the PreInterview (“Sixty Tossses”), saying “every six or seven times I think you’re going to roll the same number, again.”

For responses concerning extreme values or a range [1Aiii], some students mentioned only one end of the range, but I have grouped all responses involving range expectations or comments about either ends of the range in the same theme. The idea in this theme is that subjects give their expectations in terms of the range they expect, or they express their expectations in terms of being above a minimum or below a maximum. In adjudging supposed results as being real or made-up, the students show what they expect by how they talk about the results under consideration. For instance, they commented on how there were too few results at the ends of the distribution and they expected more, or vice versa.

[1B] Describing Why (Reasons for Expectation): The responses

themed around experiential reasoning [1Bi] were of two types. One type of response was characterized by informal, personal experiences, such as games the students had played and their recollections of the kinds of results they felt they had gotten. The other type was characterized by formal, in-class experiences, such as the activities we had done or the computer simulation we had seen. It seems clear from the results that if experiences are provided to students, they are likely to incorporate those experiences into their reasoning for *why* they hold their expectations.

A large part of students responses to *why* involved possibilities and likelihoods [1Bii]. With this theme, *what* they expected often came alongside a reason for *why*: For example, students talked about expecting to see 60 reds from the Large Jar 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. They 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.”

Proportional reasoning [1Biii] was clearly 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 Math 211 was a prerequisite for Math 212, and proportional reasoning receives a fairly in-depth treatment in Math 211, I would have expected the students to be able to reason proportionally. An interesting example of non-proportional reasoning came on the PreSurvey when one student wrote that the graphs for the Pink and Black classes (Q5b) could not be compared because the classes held different numbers of students. However, most student responses involved proportional reasoning to different degrees on different tasks, and typically the *why* of proportional reasoning went together with an average for *what* was expected.

Responses involving variation or the distribution of results [1Biv] showed up as a reason *why*, and in some ways this theme reflected how closely thinking for *what* aligned with reasons for *why*. Consider SP, who said on Q11 ("Lists" for the spinner) in the PostInterview, "you'd expect there to be greater variation and less repetition." It seems that SP is indicating that *what* she expects is to see results vary by not repeating. Her usage of variation contrasts with JM who said on Q2 ("Lists" for the Large Jar) in the PostInterview that "he's just going to have some variation". JM uses variation more as a reason. DS also uses variation as a reason when she dismissed a set of results from Q10 ("Who Cheated") on the PreInterview as unlikely, saying "there's not enough variation". Although there is clearly the potential to have a circular-sounding argument (expecting results to vary because they *should* vary), in most situations it was clear if students were

using variation as a justification for their predictions. Some students also involved the distribution in their explanations, saying for example how there should be a normal or symmetric distribution, or how a set of results had a distribution that they saw as likely. There were also responses addressing how a given distribution should be narrower or wider, or have a different shape.

[2] Displaying Variation

[2A] Producing Graphs: This dimension was necessarily influenced by students' graph sense, which is why I have listed the first theme as having to do with technical details [2Ai]. To achieve the goal of getting students to show the kind of variation they expect in a graphical form, 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. It was interesting that most students used a smooth curve in the PreSurvey to portray results for an experiment which had only 11 possible outcomes, and yet by the end of the quarter there were pictographs, dotplots, and bar charts in addition to smooth curves. 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).

When the technical details of a graph are plausible or at least understandable, the characteristics of the distribution [2Aii] can be assessed. The three characteristics that I found most salient to an understanding of

variation in the graphs of EPSTs were the center, spread, and shape of the distribution. For example, the center (or average) may be too high or low, and there may be too much or too little spread. Shape of distributions can often vary, particularly in sampling or probability situations involving small amounts of data. Similarly, the question on the PostSurvey (Data & Graphs) about average rainfall invites some interesting interpretations of what could be a reasonable shape for the distribution. Some students 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, etcetera. However, uniform distributions point to no variation: While there were a couple of rainfall graphs that implied a uniform distribution, no one had a flat shape for the sampling or probability graphs.

[2B] Evaluating and Comparing Graphs: The themes within this dimension corresponded to the characteristics of the distribution: Responses had a focus on average, spread, or shape. With the focus on average [2Bi], all three measures of mean, median, and mode were used in the collective student responses. Because these three measures were often provided off to the side of a graph in a table or box of summary statistics, it was useful to watch in the interviews how the students often went from the statistics to the graphs or vice versa, as if cross-checking to make sure the graphs and statistics matched properly. For example, by the time of the PostInterview, the students could read the median off of a boxplot, or see the mode from a histogram. But they also checked to see where the mean

was listed in the box of summary statistics, and where the mean fell on the graph in relation to the other data and measures of average. Many of the students' evaluations and comparisons of graphs focused on the measures of average, and in some cases this was the sole focus of a response. Also, I noticed a tendency to talk about the value of an average without invoking the statistical name. For example, consider GP's way of explaining why he thinks one graph shows made-up data on Q13 on the PostInterview ("Compare Graphs"):

> GP Oh yeah, like, more likely that Class A fudged the data
> DC What makes you… What persuades you that way?
> GP Because, they're all hitting above the 25 {He means the mode is at 25}, and you know, when someone wants to fudge the data, they probably think that, oh, most of the spins are going to go 25. And the other class didn't hit on any 25s {Laughs}. You know? And suddenly all these 25s are happening! {Pointing out the mode}

GP clearly has a focus on the modal value of 25, even though he never refers to the mode by name. When he said "they're all hitting above the 25," he did not mean that the data was distributed at values higher than (above) 25, he meant instead that the stack of data was literally rising highest "above" the 25 shown on the axis, thus making 25 the mode.

Another theme used in evaluating and comparing graphs was the way graphs were spread out, what the ranges were for different graphs, or how high or low the extreme values extended [2Bii]. For example, when trying to discern graphs showing real or made-up data, students readily gave their opinions about what ranges seemed reasonable, narrow, or wide, and

they expressed opinions about wanting to see more or less spread. They also used their sense of expectation for the extreme values as a guide in helping determine which graphs were more believable. The language used in comparing graphs on Q13 on the PostInterviews provides examples of the theme involving spread, ranges, and extreme values:

GP Well, Class A is a lot more compact, less range, Class B has a wider range, a lot more different variations…
JM The fact that there were no uh, nothing below I guess, 22 here on Class A, or above 28, it was just too tight.
SP This {Class B} has more variation, is more spread out, [and has] some really unexpected lows, and unexpected highs. This {Class A} is more, compact. And the range is really short {SP puts lines marking min and max on Class A}

In comparing graphs to evaluate different situations such as which city was rainier, or whether MAX wait-times were the same for the two train lines, students also considered range, spread and extremes in making their arguments.

The responses that included a focus on shape or distribution [2Biii] often included the previous two themes (focus on average and spread). By a focus on shape, I mean that students actually described the shape they either saw or wanted, often comparing shapes if there were two graphs involved. They used phrases like “a more extreme curve” as compared to “a more gentle, gradual curve,” or simply talked about wanting a “bell curve”. Responses focusing on distribution also included phrases appealing to shape, such as wanting an “even distribution” or a “symmetric distribution”, but I noticed other language that actually depicted where and how data was

distributed within the range. That is, some students talked about how data was clustered together in a certain area, or spread more on one side of the mean or the other. DS uses three terms that invoke the distribution ("scattered", "grouped", and "clustered") as she considered the two bakeries in Q9 on the PostInterview:

> DS Here {the dotplots} their plots are scattered, you know, further apart, and you're kind of – A little here, and a little there. And here {West End dotplot} you're kind of, grouped in a nice , you know, like most of them weigh right in the middle here. This one {Boxplot} you can see more…real clearly that 50 % are really clustered between 112 and a half and 115 and a half.

Especially when using boxplots, it was clear that the idea of the quartiles showing where approximately 25% of the data fell was useful to some students. In thinking about the 35 muffins for Q8 on the PostInterview, EM uses the interquartile range shown in the boxplot and then looks for the corresponding concentration of data on the histogram:

> EM 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, it does seem within, like, 112 to 115.5, it seems like that seems to be a concentration of data…

Distributional reasoning for evaluating and comparing graphs does not always go hand-in-hand with reasoning about range or average: Some students attended to the shape of the graph or the relative clustering of data without commenting on the width of the range or the location of the average.

[2C] Making Conclusions about Graphs: The first theme addresses how responses emphasized consistency or reliability [2Ci]. Some students referred to the *data* as being consistent, while other students also referred

to the *phenomenon* under study (rainfall, MAX wait-times, or muffin weights, for instance) as being consistent or reliable. I came to see that however they were using the terms, consistency or reliability were very influential concepts for some students. More importantly, responses reflecting the theme of consistency or reliability tended to have a conclusive air. For example, these three students all had similar language in discussing the rainier city on Q2 for the PostSurvey (Data & Graphs):

SA I think Columbus has more rain because the amount of rain they get every month is consistent.
RF Ohio is more consistent during all the year.
EM Columbus is more consistently rainy by looking at the boxplot.

Other students used different terminology to convey the same idea, so this theme [2Ci] is not exclusive to the terms “consistent” or “reliable.” However, many students used those very two terms, especially when discussing the MAX wait-times and the West and East End bakeries.

In making decisions in context [2Cii], some students tended to either personalize the situation or otherwise talked about the context and it’s perceived relevance. For example, when looking at which city was rainier on the PostSurvey (Data & Graphs), 11 of the 28 respondents said that Portland was rainier (despite Columbus having the higher mean and median). All of the 11 “Portland” responses included some personally relevant thought, such as SA’s comment: “I live here and it rains all the time.” A mix of context and personal preference came into play for both the MAX wait-time questions and the muffin weight questions: Students analyzed their own

tolerance for waiting for trains, or their desire for weighty muffins. On the repeated-measurement questions, such as “Car Brakes” on the PreInterview (Q5), the importance of having good brakes was a factor in deciding which graph was better for showing the data.

The “Car Brakes” question was also useful for pointing out another theme that came through in responses on the PreInterview and subsequent instruments, which was the emphasis on the level of detail and usefulness of different types of graphs [2Ciii]. By level of detail, I include the way that histograms and boxplots impart different information about a set of data, for example, and also the way that the scaling of the axes or the rounding of the data can influence a graph. In “Car Brakes”, some students preferred the graph which had an unevenly-scaled axis, expressing that the extra numbers weren’t a part of the data set and therefore didn’t need to be included on the axis. With Q7 (“Rounding: MAX”) on the PreInterview, and the corresponding question on the PostInterview (Q7 : “Rounding: Muffins”), students commented on how the different rounding schemes affected the detail of the graphs and subsequent usefulness. RL put it this way: “If it comes down to ‘I need accurate information’, [then] I want smaller increments.” Also, many responses in the PostInterviews showed an appreciation for the different types of information given by the boxplots versus bar charts, histograms, or dotplots. Some students liked the overall sense of range and subranges given by the boxplot, and others favored the in-between comparisons among data afforded

by, say, a dotplot.

[3] Interpreting Variation

[3A] Defining Variation: The first theme of definitions and descriptions [3Ai] was informed by one of the very earliest questions on the PreSurvey, asking students what "variation" meant to them. In addition to attaching a mathematical connection to the term "variation", responses also reflected the idea of variation as involving differences or changes, or having choices or options. Some students also referred to the degree of difference or change.

I made it explicit in explaining the research project to the class that I was studying conceptions of variation, and through the probability and statistics part of the course, Steve and I freely used the term "variation" and associated terms. Therefore it is no surprise that students also specifically used the term "variation" in all the contexts of the research instruments. What emerged was an implicit picture of what variation came to mean to the students. That is, I did not ask at the end of the class "What does variation mean to you now?" but instead I looked back on all the ways in which variation was reflected in the responses to come up with added definition and description for the term "variation".

Two important uses of the term are to describe the range and also the distribution of data within the range. For example, one student referred to seeing the variation in a graph in terms of the range shown on a boxplot, and the same student also mentioned seeing the variation (for the same set of data) in the way there were many different categories and frequencies shown

by the histogram for the data. The two uses (as a range and for the distribution within the range) match well with the earlier ideas of variation as a *degree* of difference as well as *having* differences. Other descriptors I found for variation included the way that data was clustered, spread, concentrated, grouped, and distributed.

The second theme of examples of variation [3Aii] was asked also on the PreSurvey, again with a few students offering math-related examples. Other main types of examples reflected characteristics of natural or personal characteristics, showing what came to their minds in thinking of variation at the outset of the quarter. The dominant commonality among most of the examples given by students on the PreSurvey was that the examples hinged on having differences, such as “chords in a song” or “the shapes of rocks.” Throughout the subsequent research instruments, students implicitly affirmed examples of variation via their responses. For instance, students provided choices on “Six Trials” 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 showed 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.

[3B] Causes of Variation: From the surveys and interviews, there were four questions in which I explicitly asked students what they thought might be

some causes of variation for these situations: Rainfall patterns between Columbus and Portland, traffic deaths between the South and the NorthEast regions of America, MAX wait-times, and muffin weights. The last two situations (MAX wait-times and muffin weights) involved repeated-measurement scenarios. Beyond the four questions asking directly for causes of variation, students also volunteered possible causes on additional or related types of situations, such as sampling from the Small and Large Jar and also using the spinner. I found that students gave myriad causes, many of which seem to have two main themes: Naturally occurring causes and physically deliberate causes.

The theme of naturally occurring causes [3Bi] includes, for example, randomness as reason for variation in sampling and probability situations. The theme [3Bi] also encompasses different reasons for weather variation. I also include in this theme reasons such as different speed limits, drivers' age requirements, or road conditions (for the traffic deaths situation), because those are normal, routine reasons which might account for variation in the data. Also, many reasons for different MAX wait-times - 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, having crumbs fall off a muffin as it gets repeatedly weighed seems a normal occurrence.

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 scenario, an example of a physically deliberate cause [3Bii]. There were some other causes falling into this theme [3Bii], most notably having to do with the sampling and probability environments. For 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, the responses emphasized the way that the hand might grab the candies, for example, thus making the situation seem as if the person doing the sampling was causing the variation. With the spinner, more than the other random devices (the coin or the die), 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.

[3C] Effects of Variation: For the theme of direct effects [3Ci], I have included responses where it seemed to me that the student had made a direct connection of some kind between variability and different outcomes. Students talked, for instance, about how what gets predicted by probability and what happens in reality do not always match. Whether or not students see variation as a cause of how probability predictions do not always match (on an outcome-by-outcome basis) with reality, the observation centered on “probability versus reality” is reasonably astute. There are likely to be different outcomes in a probability experiment, even with an underlying theoretical ratio, and the presence of differences can be seen as an effect of the variation

inherent in the experiment. Similarly, in referring to how results won't be "perfect" every time (whether in a sampling or probability situation), the idea was that this portrayal of results as not "perfect" was a direct effect of variation. Students also talked about the likelihood of different events, such as getting a certain weight of muffin, or waiting a certain amount of time for the MAX train, and again my sense was that students viewed the different likelihoods as the result of the presence of variation in the data.

Whereas the direct effects described above are at least somewhat plausible (real data *does* often differ from predicted value, and results *aren't* "perfect" every time), the indirect effects [3Cii] include responses that are either less plausible or that reflect poorer understanding of variability. For example, students talked about how results "Could Be Anything" ("Anything Can Happen"), and thus "You Can Never Know". On the surface, the attitudes of "Anything Can Happen" and "You Can Never Know" reflect a degree of logic in terms of what is possible or guaranteed. If an event is in a sample space, for example, that event can happen. Having a priori knowledge of future outcomes is an unreasonable expectation. However, the "Anything Can Happen" attitude was often expressed as if there were no concomitant understanding of how unlikely some events were. The "You Can Never Know" attitude seemed to imply that a person could never make a reasonable claim. Thus, the two attitudes often went with a difficulty in making choices over what results might occur, or a lack of confidence in determining real data from made-up results. I do not claim that the students expressing the above

attitudes see their perceptions as an indirect effect of variation. Rather, I wonder what causes them to have those points of view. It may be a lack of understanding about variation. It may be akin to the outcome approach, as discussed in the case of SP in Chapter Six. It clearly has something to do with uncertainty. In the final analysis, I suggest that *because* the situations are inherently variable (and because students may poorly understand what constitutes reasonable expectations of variation), *therefore* students express difficulty making predictions, thinking that "Anything Can Happen" and "You Can Never Know."

[3D] Influencing Expectations and Variation: The first theme of sample and population size [3Di] came through as I considered responses that focused on the numbers of candies in the Small and Large Jars. One type of response was akin to what other researchers termed "Additive Reasoning", meaning that the focus was not on the ratio but on the sheer numbers of candies in the jar. Here are two examples from the PreInterview:

EM [Q1] Well, my thinking is that you're going to get close to 6, since there's more reds than yellows, you're likely to pull, you know, at least half will be red.

SP [Q1] Because there's more red than yellow in the jar… you're going to get maybe… maybe more red.

EM and SP also gave proportional reasoning on other parts of their responses to Q1 on the PreInterview, but without knowing that they might be thinking proportionally, 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. Interestingly, EM and SP had similar responses on the PostInterview:

> EM [Q1] Because, there're 600 reds , 400 yellows, so there's more of a chance of pulling reds because there's more reds in the jar.
> SP [Q1] There's also that 400 yellow in there, so… The likelihood of getting all – like, of 600 red and 400 yellow, the likelihood of just getting the 60 out of 100, which is like the perfect ratio or whatever – is very unlikely.

It is not clear from the above excerpts if EM and SP perceive the chances of red to have increased because of the increase in numbers or not. However, it does seem that the sheer size of the jars and samples is influential in their (and other students) thinking. Also, when comparing the graphs on Q5 ("Small & Large") in the PostInterview, the relative size of the two classes' samples and populations was relevant to some responses:

> GP The thing is, this {Class A: Small Jar} is a wider range, and this {Class B: Large Jar} should have the wider range. Yeah, because you're grabbing from a larger group, and you know, should have a larger range than that {Small Jar}, the smaller container.
> RL I might have expected a more awkward , I guess, configuration or graph in Class A, with the small sample size. This {Class A: Small Jar} points to the theoretical distribution pretty accurately, and with such a small sample size, I might not expect to see that.

RL's response may point to some understanding of how smaller samples vary more than larger samples, while GP's response suggests that larger samples should have larger ranges than smaller samples. Whether reasonable or not, the responses within this theme [3Di] showed a connection between the numbers with a sample or population and expected results.

The last theme, concerning number of trials [3Dii], had several

characteristics. One characteristic was that more trials would yield more variation, and in this sense variation was used as a synonym for range. In other words, doing more trials would extend the range in both directions. Another way that students had of expressing this view was to say that more trials gave more chances to get extreme values, or as JX put it on the PostSurvey (Probability): “The more sets done, the more likely you will get less likely results.” A second characteristic was that more trials 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 that suggested, for example, “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” (JM on Q13 in the PostInterview, “Compare Graphs”). A third characteristic was that more trials gave a better picture of the underlying distribution, and fewer trials gave an impoverished picture. Here is an example of a response showing this third characteristic, as DS considers two graphs on Q4b (“Compare 30 & 300”) in the PostInterview:

> DS 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.

In the PreInterview and on the PreSurvey, there were also responses for the fourth characteristic, which was that the number of trials had no effect on

the probability associated with individual trials. While true, the responses were often used to justify an expectation of no variation in results from repeated trials.

Cross-Case Analysis

The emergent framework proved useful for looking at individual conceptions of variation, as was described in the second part of Chapter Six when considering the six cases' responses to a common subset of PostInterview questions. I used the framework in another way that I'll now describe, which helped me see similarities among and differences between cases.

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 emergent 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 ("Lists" for the Small Jar) on the PreInterview. The interview script had me ask for the 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 65 below to show me how the different cases' responses fit into the framework:

CodeFrame for PreInterview Q2
(Cross-Case Analysis)

Framework	Description Within Theme	Subject (Case)					
1Ai	Should be on Both Sides of Exp. Val.	DS	EM		JM	RL	SP
1Ai	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				
3Cii	Difficulty in Making a Choice			GP			
3Cii	Anything Can Happen				JM		
3Dii	Expected Value as an Average		EM				
3Dii	More Trials = More Variation	DS					

Table 65

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.

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 65 above, 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 1 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]. What I will summarize in this section are 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 1 has 15 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 6 cases for certain rows, which I

refer to as “Match 6 Rows”. The “Match 6 Rows” are what I’ll be summarizing next. There were also “Match 5 Rows”, meaning that exactly five of the six cases were coded along that row (as in the three rows of Table 1). Table 66 below shows how many CodeFrames I had in each interview, how many rows were in each CodeFrame, and also how many Match 5 or Match 6 Rows were in each CodeFrame.

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

Table 66

I chose to show Match 5 or Match 6 Rows in Table 2 simply 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. It is the fourteen Match 6 Rows in the PreInterview and the eighteen Match 6 Rows in the PostInterview that are summarized next, because they represent the most agreement.

PreInterview

Knowing from Table 66 that fourteen rows from different CodeFrames have all six cases' responses matched does not tell us what those rows are: That is, we cannot see from Table 66 which part of the framework 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 using the framework to re-organize the fourteen Match 6 Rows from the PreInterview, Table 67 below shows exactly what the fourteen rows represented:

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	Q11a
1Aii	Their own choices are all different	Q11b
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	Q4a
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 67

Although not organized by question number, Table 67 does show precisely what I wanted to know: 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 67 in terms of the emergent 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 Q11a and Q11b ("Repeat Trial" and "Six Trials" with the die). 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].

In *displaying* variation, everyone had a focus of attention on the averages [2Bi] shown for the graphs in Q7 ("Rounding: MAX") and Q8 ("MAX Wait-Times"). I was encouraged to see that all subjects commented on the short range depicted in the graph on Q3 ("Fake: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 Q4a ("Real:300"), however, 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 68 on the next page. 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 responses had to do with *expecting* variation, and for Q1a and Q10a ("One Trial" at the Large Jar and spinner, respectively) each of the cases did not just list the expected value for their prediction, and instead they gave answers indicating an appreciation for variation. They also all favored list (ii) on Q2 ("Lists" for the Large Jar), which was the most reasonable choice. The theme concerning the expected value [1Ai] was

subscribed to by all cases in Q11 (“Lists” for the spinner) and in Q10b (“Compare Trials” for the spinner), with responses indicating that responses should be close to the expected value.

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

Table 68

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 (“Compare

Graphs") 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 ("Rounding: 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: "Reasons: 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].

Discussion

Finding out where within the framework there was detailed agreement among the cases was one of the main reasons for using the CodeFrames as

described above. There are many other potential uses, such as finding where there were more unique responses, or how cases compared by gender. The CodeFrame approach essentially represented a microanalysis of the interview data, because the unit of analysis was each line of text in the interview transcript. That is, I was able to take every part of a case's response and evaluate that part to see what dimension, theme, or characteristic of a theme was reflected. As explained earlier, some themes, such as "Concerning Expected Value," have many characteristics: Results should be close to expected value, results should fall both above and below the expected value, etcetera. The cross-case analysis presented in this chapter shows the lens of the framework focused as finely as I have been able. Although matched responses did not require a word-for-word correspondence among cases, in order for there to be agreement the cases needed to express the same reasoning within the given dimension.

Using the CodeFrames as I've done in the cross-case analysis shows some overall trends, most notably the closer agreement in *expecting* variation. For example, whereas on the PreInterview there were many predictions for "One Trial" questions that were simply the expected value, on the PostInterview there were ranges given for predictions, or values that were explained as being near to the expected value. Also, on the PreInterview, some cases did not seem influenced much by proportional reasoning, while other seemed overwhelmingly influenced by theoretical predictions. On the

PostInterview they all used proportional reasoning but did not take the view that such reasoning offers point estimates that always occur. As DS pointed out in Q12 (“Compare Comments”) on the PostInterview, “there’s no guarantee that there’s.. That you’re going to get exactly 25 out of 50. Because, the spinner is… Um, randomly landing.” There were also some uniformly reasonable conclusions made regarding *displays* of variation, as well as attention given to average and range in evaluating and comparing graphs. Finally, in the PostInterview there was total agreement about plausible *interpretations* of variation, namely the cause of variation in Q6 and the influence of more trials on results in Q12.

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. I’ll briefly highlight some shifts among the six cases before turning to the final summary of results in response to the research questions.

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 (“Fake: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 earlier talked in the quarter about results not being “perfect”, by 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 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.

Finally, whereas RL was highly motivated by theoretical expectations at the outset of Math 212, 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.

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.

Summary of Results

The impression of improved reasoning for the six cases has support from cross-case analysis as well as from the survey results presented in Chapter 5. The survey data came from the entire class, and the analysis of results gave support for overall improvement in reasoning on all three aspects of *expecting*, *displaying,* and *interpreting*. For example, thinking of the class

as an entire case, choices related to *expecting* on “One Trial” and “Six Trials” improved in the probability context from the PreSurvey to the PostSurvey. Also, better graphs were produced on the PostSurvey (Probability) than on the PreSurvey, showing improvement in *displaying*. Finally, related to *interpreting*, reasons for having an expanded range were better for similar questions from the PreSurvey to the PostSurvey (Probability). Altogether, it seems clear that the classroom interventions – and likely the interaction with the research instruments themselves – affected the EPSTs conceptions of variation. Several students wrote on the PostSurveys or talked in the PostInterview about what they had seen in class, offering their experiences as justification for their thinking.

However, as emphasized previously, the purpose of this research has not been to prove effects of classroom interventions. The goal of the research, as illustrated by the questions asked at the outset and kept as a guide throughout the entire design and implementation, has been to describe and characterize the conceptions of variation held by EPSTs. It is useful to point out shifts in thinking, and to note the impact of classroom interventions, because research specifically designed as a teaching experiment would be the next step for future research (to be discussed in the following section). That is, I now have some better ideas of good tasks and class activities that show promise for addressing all the aspects of the framework. Prior to this research, it was an open question as to exactly what

aspects would be involved in EPSTs' understanding of variation, or how those aspects might be described.

Thus, it is the emergent framework, grounded in the survey and interview data, which addresses the research questions. 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, it remained an open question at the outset of this research whether or not EPSTs did think along the lines suggested by the initial framework posited at the end of Chapter Three. By the conclusion of the research, a deeper exploration of EPSTs conceptions about variation showed not only how those conceptions mapped into the framework, but also that framework itself gained richness in detail.

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, as described so far, are useful for looking at EPSTs' conceptions of variation, but are not guaranteed to easily characterize all kinds of 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 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 refinements in the next section.

Another possible limitation of the research concerns the class environment. The culture of the Math 211 and Math 212 classes at PSU 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 Math 211 at PSU. 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, and a third area concerns the curriculum with teacher preparation. I'll briefly discuss these three areas before including some additional implications for current practices in teacher training.

In the first area, 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. What became problematic was when the students had alternate meanings within the same

response, and some new tasks or new lines of questioning could be designed to address these problematic situations. Also, using some of the survey items on a large scale with preservice teachers at several universities would accomplish two useful purposes. The first purpose is that the overall utility of the framework could be tested on a stronger quantitative basis than was offered in this research, and it could then be seen how generalizable the application of the framework is. The second purpose is that 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 researchers have a sense of what are the important aspects to be looking at. This research provides a critical first step towards identifying the important aspects and what comprises those aspects.

For the second area, previous research has looked at or is looking at conceptions of variation held by elementary, middle, and high school students. This research looks at prospective teachers of students, so 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.

Regarding the third area of the curriculum for teacher preparation programs, a study designed specifically as a teaching experiment would be appropriate, now that this research has pointed out relevant aspects to focus upon as well. This research has also laid out some useful interventions to consider, but to actually measure effectiveness in a classroom setting seems to require additional research that aims more at the teaching and learning within the class. Steve is a seasoned Math 212 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 Math 212 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.

However, this research already provides ample 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 of *expecting*, *displaying,* and *interpreting* variation. They need these opportunities with 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 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 *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.