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.