CHAPTER FOUR

Results and Analysis

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

Evolving Framework

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

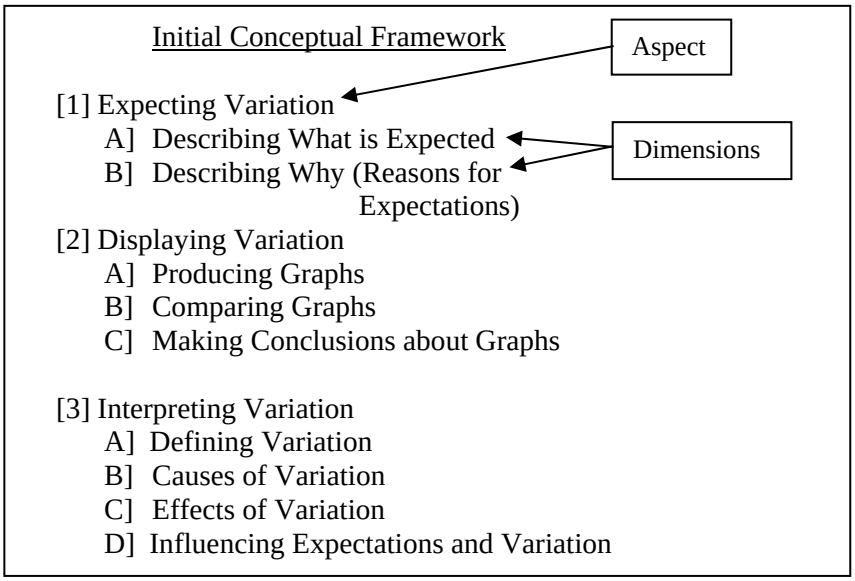

Figure 12 – Initial Conceptual Framework

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

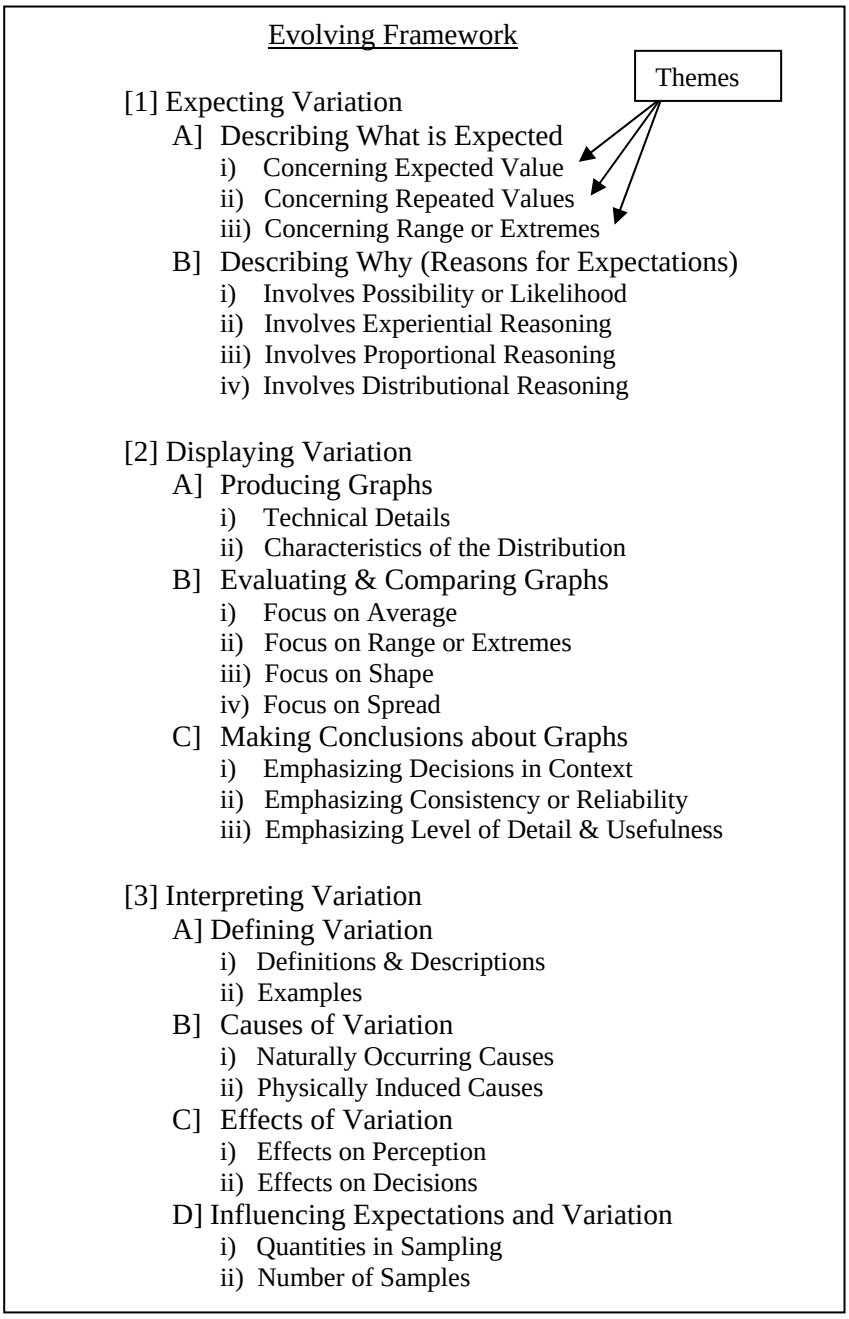

Figure 13 – Evolving Framework

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

[1] Expecting Variation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

expected value would occur twice.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

between informal and formal experience as a way to group the responses I found for this theme.

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

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

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

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

Here were some of the reasons offered by subjects for their predictions on the Sampling PostSurvey:

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

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

iii) Involves Proportional Reasoning – Proportional reasoning was a part of almost every student's explanation at some point in their individual responses. The variety of ways they collectively had to explain included ratios, decimals, odds, and

fractions. Since MET 1 was a prerequisite for MET 2, and proportional reasoning receives a fairly in-depth treatment in MET 1, I had expected the students to be able to reason proportionally. Some sample responses from PreSurvey Q1a (reasons for the predicted results for one sample of size 10) show the diversity that subjects had in expressing their proportional reasoning:

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

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

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

expected.

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

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

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

the way that data clusters close to a center, or spread from the mean, or is concentrated at various intervals within the range.

Distributional reasoning can therefore encompass several different characteristics, and some subjects incorporated elements of distributional reasoning into their responses to varying degrees. I'll discuss some of the better responses exemplifying this important theme, choosing examples from the PreSurvey and each of the PostSurveys. In PreSurvey Q2, students were asked to predict a range for six samples, then a range for thirty samples, and then to offer an explanation. MA listed a range of 3 to 9 for both six and for thirty samples, and wrote

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

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

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

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

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

SC wrote about the average and how data was distributed about that value. She also mentioned the shapes of the distributions, and in describing her "restacking" she exhibited a powerful sense of distributional reasoning for this question.

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

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

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

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

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

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

following four responses:

JB: The highest number of people draw from 51-60. If graphed, the graph would be symmetrical.
MM: Seems like there would be a concentration here [in subrange 41-70] and then the others would be the outliers or less likely pulls
SC: Because they create a "picture" of data that peaks around 60 and clusters around that mark, diminishing as it moves to the extremes of 100 and 0.
SA: Because 51-70 is closest to the mean, so those will happen the most times. As the numbers get farther away from the mean, they will happen less.

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

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

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

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

I gave more examples for this theme than for previous themes because of its importance in revealing thinking about variation. The individual characteristics within the theme – centers, range, shape, and spread – are at least as important as the mix of

these characteristics within a response and the language that subjects use. Many of the examples given for this theme are lengthier precisely because some subjects were relating different elements of the distribution together. Subjects also may lack conventional terms such as "standard deviation", but they are still capable of conveying a sense of reasoning about the distribution of data.

[2] Displaying Variation

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

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

Types of graphs that subjects used included smooth curves, bar charts, dot plots, scatterplots, pictographs, and straight lines. Sometimes they labeled their axes and put appropriate scales and sometimes they did not. To illustrate some of the different types of graphs, first consider PreSurvey Q4, which asked for a graph showing predicted results of 50 samples of size 10 taken from the Small Jar. GP drew

the skewed bell shown in Figure 14. I had provided labeled axes on the PreSurvey, and placed a scale on the horizontal axis. In GP's case and several other subjects, no scale was given for the vertical axis, making it hard to tell how plausible his graph was.

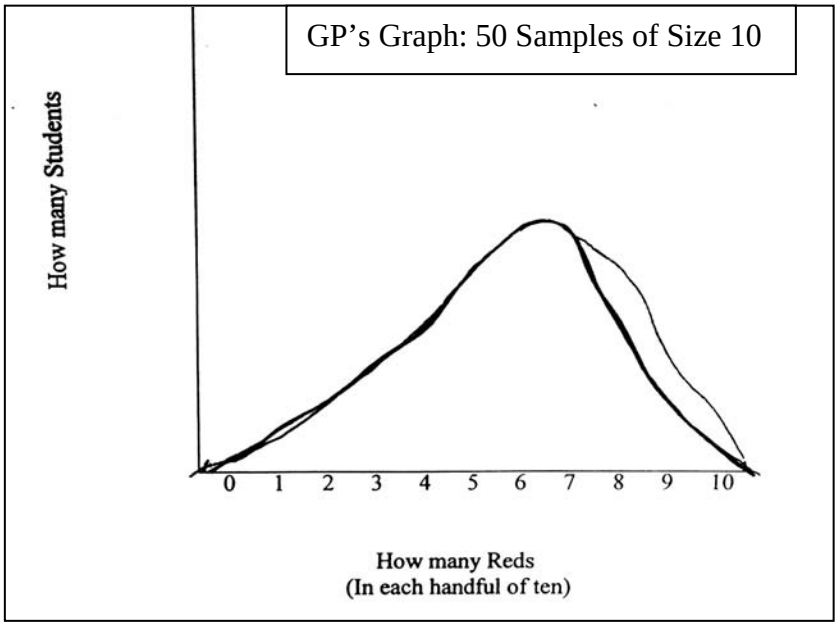

Figure 14 – GP's Response to PreSurvey Q4

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

By the time of the PostSurveys, we had practiced making several different types of graphs in class. I'll next share a response to Data & Graphs PostSurvey Q1c. The question asked for a graph showing how many inches of rain Columbus, Ohio might get for each day in June, assuming that the average monthly rainfall for June was 4 inches. Figure 15 shows BP's graph. Again, I had pre-labeled both axes, and I also had subdivided the horizontal axis to show marks for each day.

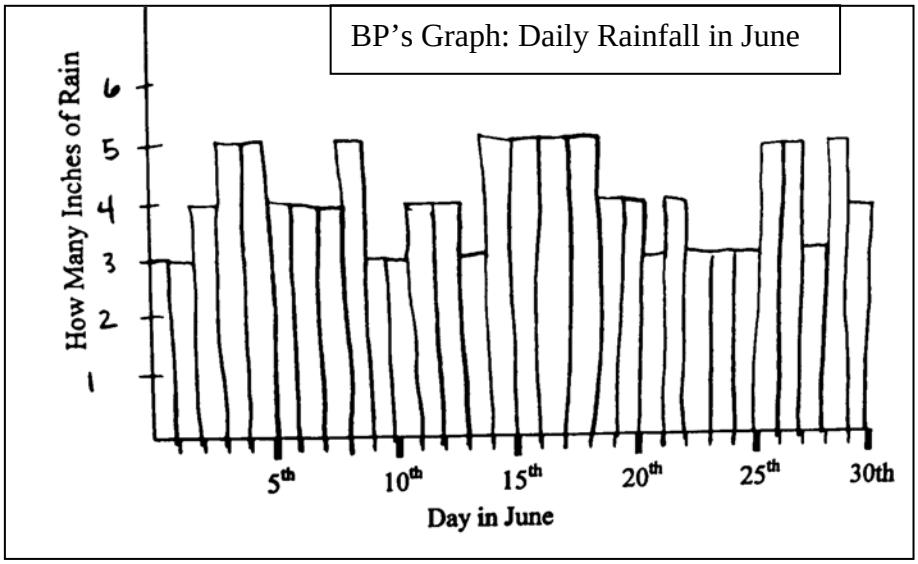

Figure 15 – BP's Response to Data & Graphs PostSurvey Q1c

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

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

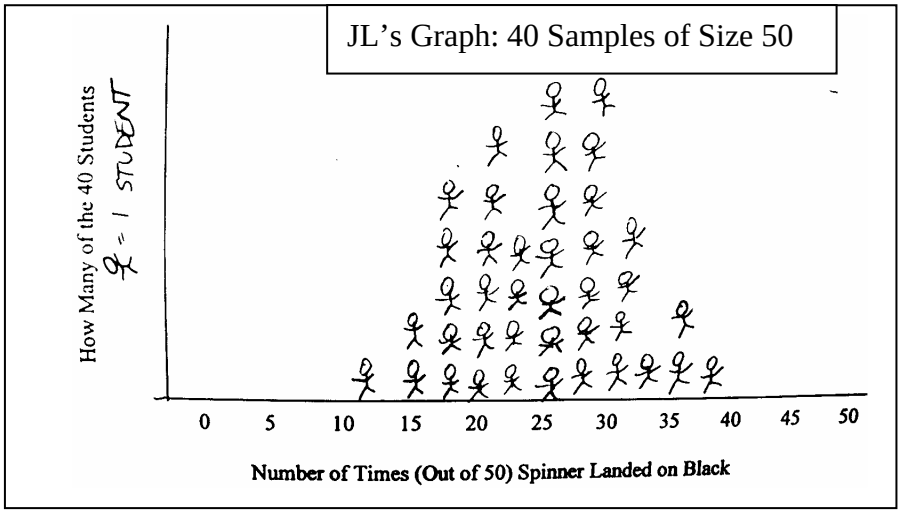

Figure 16 – JL's Response to Probability PostSurvey Q3

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

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

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

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

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

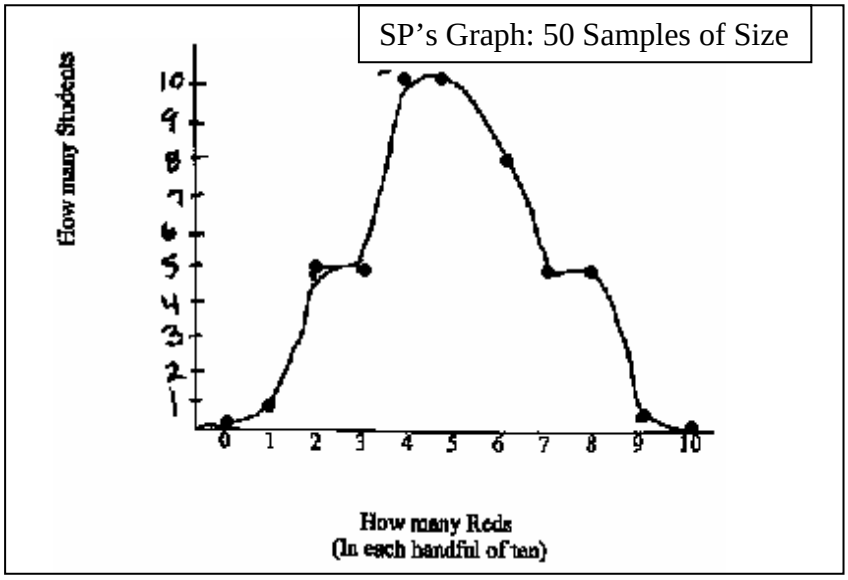

Figure 17 – SP's Response to PreSurvey Q4

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

had an unrealistically low expectation concerning the upper end of the distribution.

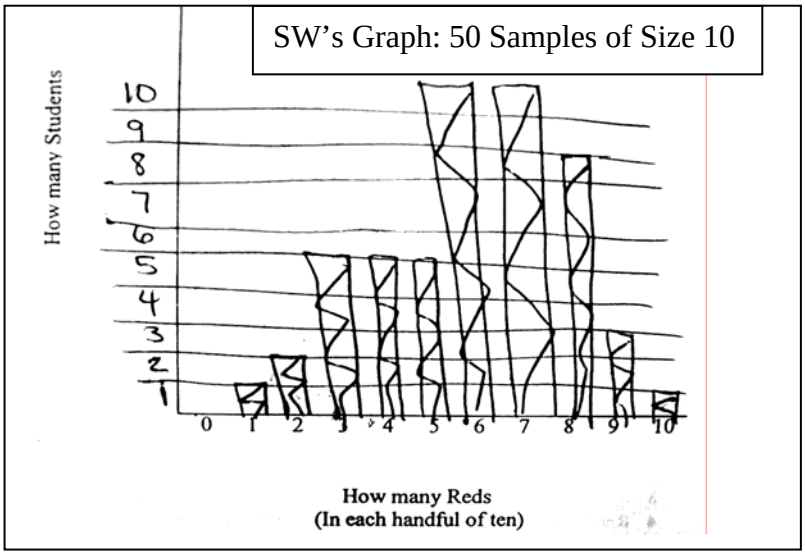

Figure 18 – SW's Response to PreSurvey Q4

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

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

add up to 4" was common for many students who obtained the correct average.

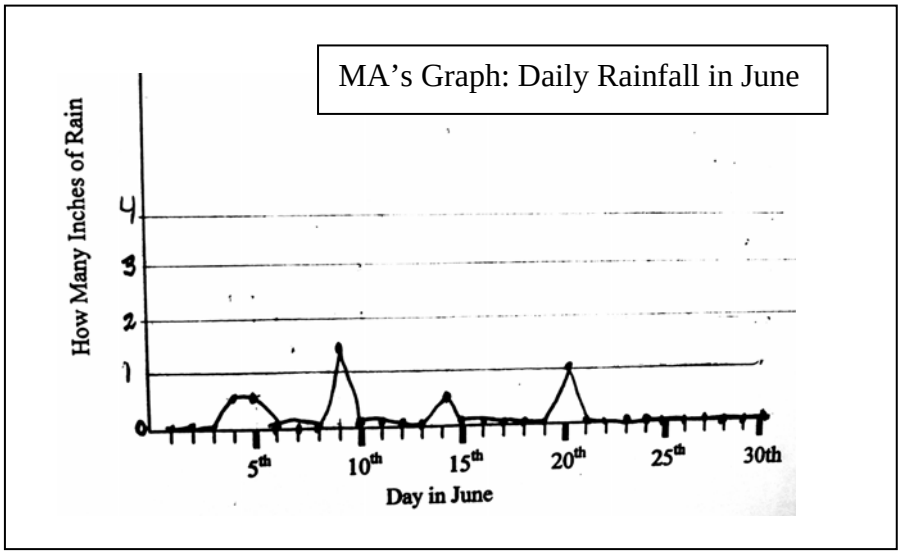

Figure 19 – MA's Response to Data & Graphs PostSurvey Q1c

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

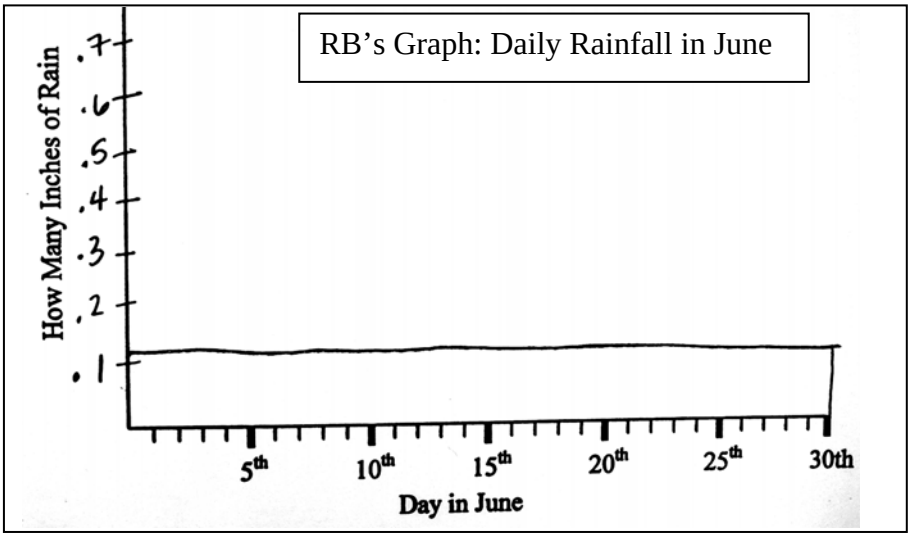

Figure 20 – RB's Response to Data & Graphs PostSurvey Q1c

RB's horizontal line conveyed the notion of absolutely no variation, and a steady flow of rain all through the month of June. Another subject, RL, also used a straight line but had the line increasing from left to right, saying that the "graph is inclined because [the] average for July is greater than 4". Most subjects showed some day-to-day variation in their rainfall distributions, even though most did not correctly determine the daily average.

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

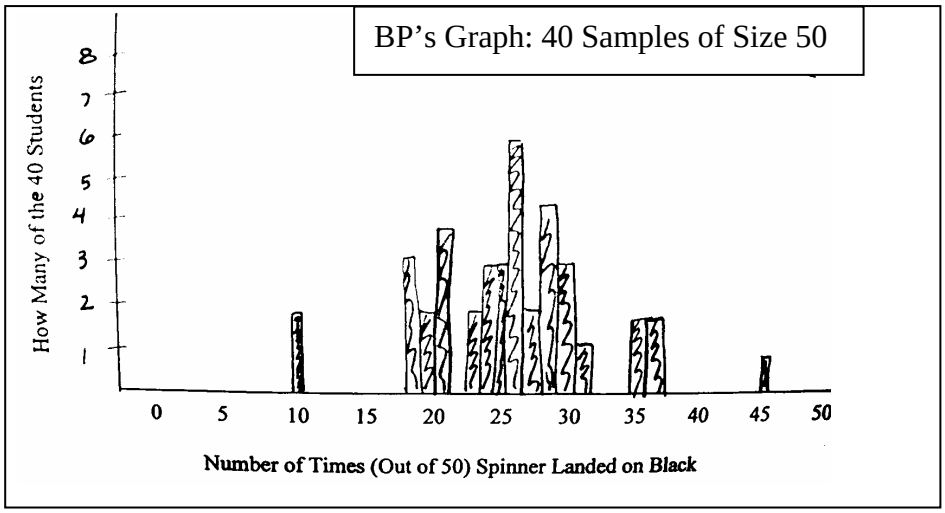

Figure 21 – BP's Response to Probability PostSurvey Q3

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

from the mode on both sides.

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

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

i) Focus on Average – On many of the tasks, a box of summary statistics was provided, listing the mean, median and mode for the data. The box was usually put in

close proximity to the graphs, and some subjects were so influenced by those numbers that they admitted to not even paying much attention to the graphs. All they needed to know was what the measures of center were.

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

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

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

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

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

For this question, boxplots and bar charts were provided to show the sets of data for the two cities, and summary statistics were also provided. From the boxplots, subjects

could see that Columbus' median was higher than Portland's, but to compare the means the subjects referred to the box of summary statistics.

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

DS: Because... here [On the boxplot]... your median is right at, like, 113 and a half, so … And here [On the histogram] your mode is at 113

GP: Well, the median [He points to summary statistics] is 113.5 grams.

RL: Well, I'm looking at the mean, [He points to summary statistics] and I'm looking at the mode, in this case, which really stands out...

JM: Well, I look at the median as it's written out [In the summary box], and I also look at the amount of muffins at 113 [On the histogram], which is… you know, the mode is actually 113 too, and here [On the boxplot] it's also the median …

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

ii) Focus on Range or Extreme Values - PreSurvey Q6 invited a comparison of student heights at two different schools, and the question elicited many responses that focused on range. For the two bar charts shown in Q6, School A had a wider range of student heights than School B. However, in School B, the heights of the adjacent bars varied up and down more frequently than the smooth rise and fall of the bars in

School A. Some subjects discussed ranges without using numerical descriptions, making it difficult at times to tell if the subjects were referring to the range of bar heights or to the range of student heights:

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

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

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

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

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

DS: Because they have fewer on the ends [She points to the ends on Class A]
SP: These lower numbers are a little surprising, for both of them.
EM: In my opinion, you might get a few more 9s, and maybe a 10.
GP: Well, I think it's pretty hard to get these 2 and 9s… I mean, really hard. See, the thing is, this [Class A: Small Jar] is a wider range, it seems like, than this [Class B: Large Jar]

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

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

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

GP: Top of the pyramid is 6 or the most probable and it just cascades down.
MA: I see the result as a bell curve, since there is greater chance of getting more red than yellow, but getting ALL red is not likely either.

MG: Because in random sampling, shouldn't they fall into a bell curve?
RL: A bell curve represents the most likely scenario – the extremes aren't seen often, the average is seen the most often.

In class, I also heard other students echo GP's use of "pyramid" to refer to an inverted-"V" shape, which also connoted the idea of a bell curve to many.

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

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

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

Sometimes just certain elements of the shape stood out to subjects, such as the heights of the tallest columns in a bar chart or histogram. For example, on the Data & Graphs PostSurvey Q1, the graph for Portland's normal monthly rainfall has tall bars

for the winter months and shorter bars for the summer months (denoting heavier and lighter rainfall). Several students took note of the shape for the Portland data by emphasizing the dominant winter months:

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

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

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

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

DS: Well, Graph 2 you can see that... it's kind of spread out, and it's not as, at a glance it's not as, like, you kind of your eyes go "wooaahh" [She dramatically waves hand from one side of table to the other]

EM: Graph 2, I don't know. Graph 2 to me is too spread out, so I'm... I like seeing the Xs next to each other, so I can compare them easier, whereas Graph 2 is kind of spread out and I can't really read.

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

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

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

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

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

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

When CM wrote about when Columbus "gets most of its rain", she is used a naive form of spread since she doesn't mention how much is "most". Later in the research,

subjects made references to percentages, such as the upper 75% of the data, to help quantify where they saw data clustered.

The Interquartile Range (IQR) on boxplots applies the middle 50% of the data, and is one measure of spread. In Data & Graphs PostSurvey Q1bii, subjects referred to the IQR in discussing which city was rainier, Columbus or Portland:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

I'll start with examples from the car brakes situation on PreInterview Q5. In the interview script, I introduced the term "consistent" as a part of a subquestion: "If the engineer wanted to suggest that the car was fairly consistent in its braking power, which graph would you suggest she use, and why?" Admittedly, this phrasing plants

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

numbers".

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

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

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

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

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

axis (as in Graph 2).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[3] Interpreting Variation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Another characteristic concerned people or personal characteristics:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

theme is that someone or something acts in a purposeful way which introduces variation into a situation as a result of the act.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

drawing samples from the Large Jar, there were other responses similar to SP's:

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

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

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

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

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

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

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

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

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

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

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

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

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

by *proportion*. For instance, SW explicitly mentions how "there are 20 more reds", showing an additive strategy.

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

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

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

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

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

Thus, the sheer size of the samples and populations featured prominently in subjects' reasoning about both Small and Large Jars. There are some useful ways in which moving from a Small to Large Jar does influence the distribution, but I hypothesize that many subjects only think in terms of greater numbers leading to higher or lower likelihood. For instance, they may think that trying to draw a red marble from a 60 Red/ 40 Yellow mix is an easier (likelier) task than trying to draw a red marble from a 600 Red/ 400 Yellow mix.

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

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

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

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

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

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

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

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

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

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

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

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

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

Summary

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

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

Individual Cases

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

--- Continued in Ch4_PtB ---