CHAPTER FIVE

Survey Results and Analysis

Introduction

In the first part of this chapter, the results from the classwide data are summarized. The goal in using the classwide data is to validate the aspects and dimensions of conceptual framework posited earlier. That is, I wanted to see if the framework was useful in helping look at and organize the responses from the surveys, and also I wanted to see what themes developed for each dimension. Recall that the primary aspects were *Expecting*, *Displaying*, and *Interpreting*, with the corresponding dimensions listed below each aspect of the framework:

[1] Expecting Variation
  A] Describing What is Expected
  B] Describing Why (Reasons for Expectations)
[2] Displaying Variation
  A] Producing Graphs
  B] Comparing Graphs
  C] Making Conclusions about Graphs
[3] Interpreting Variation
  A] Defining Variation
  B] Causes of Variation
  C] Effects of Variation
  D] Influencing Expectations and Variation

For example, *Producing Graphs* above is a dimension of the aspect of *Displaying Variation.* The classwide data added details to what was involved in *Producing Graphs*, and I will refer to categories of common responses as themes of the dimensions. These categories began to emerge with the PreSurvey results, which are next summarized. After the PreSurvey, some

results from the other written documents completed by the class are presented.

PreSurvey

The first section of the PreSurvey, consisting of just the first page, contained the background questions discussed in the previous chapter. Also on the first section, Question 5a asked: "What does the word 'variation' mean to you?" There were four categories that came out of the responses, and it was possible for different elements of a student's response to fall into different categories. Table 9 descriptively names the categories and also shows how many responses got coded at each category.

Meaning of the word "Variation"		
Category Description	Number of Responses	Sample Response
[A] Having Differences or Changes	18	SP: The differences between things in a group DS: Changes over time
[B] Degree of Difference or Change	8	CM: Degree to which something is different JL: The degree by which a number can change
[C] Having Choices or Options	3	RF: When you have more than one option or result SL: Alternatives
[D] Uses Math Language	6	BP: How far something deviates from the average LW: Distance from the norm

Table 9

The category which captured parts of the most number of responses was described as "Having Differences or Changes". What sets this category apart from the others was a focus on the simple presence of differences or changes, whereas in "Degree of Difference or Change" the emphasis was on how much of a difference or change there was. Certainly "Having Choices or Options" implies that there are differences in those choices but the emphasis seemed to

be more on decision making. That is, you have alternatives and you then decide what to do, or you have different results based on what alternative you choose. Thus, this category goes beyond the mere presence of differences. Lastly, in "Uses Math Language", responses were characterized not by a relation to a personal or real-life context, but only by a connection to mathematics. As one student wrote, "I think it has something to do with equations."

The second part of the question, 5b, asked subjects to "Give an example of something that varies," and the responses are summarized in Table 10 below. Some students gave no response, while others gave more than one example in their response.

Examples of what "Varies"		
Category Description	Number of Responses	Sample Responses
[A] Relating to Math	3	JL: The number 10 varies from 10.1-10.4
[B] Natural Phenomena	9	JM: The shapes of rocks. Snowflakes
[C] Relating to People	11	SP: Weight, height, hair color of a group
[D] Relating to Finances	3	BP: The stock market
[E] Relating to Time	3	SC: Time of day
[F] Relating to Music	2	JX: Chords in a song
[G] Relating to Food	2	SL: How many chocolate chips are in a cookie

Table 10

Taken together, both parts of question 5 on the first section of the PreSurvey added detail to the *Interpreting* aspect. Specifically, I saw that the dimension of *Defining Variation* could hold two distinct themes: Definitions or meaning of variation, as given in Table 9, and also examples of variation, as given in

Table 10. It can be seen that the dominant thinking for definitions and examples of variation was not in a strictly mathematical sense, but more in terms of general differences having to do with real-world phenomena. Since the treatment of the definition of variation for my research as described in the first chapter is quite broad, I was curious to mine other ideas that came through the subsequent survey and interview questions about how these EPSTs defined variation. In the last chapter, I show results from doing a text search on all responses (from the surveys and interviews) for a set of terms that arose in the data further describing data.

In looking at the responses on the second section of the PreSurvey, two approaches were taken in connection with one another. One approach assessed responses according to a coding scheme, and the second approach used my framework to look at the categories from the coding scheme. Part of the coding scheme for the PreSurvey was used with identical questions on surveys for the NSF project involving middle and high school students. I developed the other part of the coding scheme to address similar questions that were a part of the PreSurvey but not asked in the NSF project. The primary purpose of the coding scheme is to look for trends in responses and to group those responses together accordingly, and the reason I chose to use and extend the NSF coding scheme is because I had been a part of its development and had already used it extensively. It gave me a way to see some initial categories of responses, after which I looked at the categories with an eye for how those categories might fit into or inform my framework.

Two colleagues (Matt & Kate) who still were involved with the NSF project took my PreSurvey data and also coded the responses, which strengthened the final fit of responses to categories. With 19 different parts of questions coded for 27 students, there were a total of 486 codes assigned. When I finished checking our codes, there were 442 codes that matched, for a 91% agreement. The questions and the codes, as well as sample responses, are next provided. I'll organize the second section of the PreSurvey by the main context the questions are related to, and the first context is sampling.

Sampling

The first four questions of the second section of the PreSurvey all involved trials that consisted of handfuls of ten candies taken from a jar containing 60 red and 40 yellow candies. The number of red candies in the handful of ten are counted, and then the candies get returned to the jar and remixed. Table 11 below summarizes the first three questions, which collectively related to the aspects of *expecting* and *interpreting*.

PreSurvey Questions #1-3		
Name of Task	Question Number	Description of Task
One Trial	Q1ai	You do one trial. How many reds do you think you might get?
	Q1aii	Why do you think this?
Several Trials	Q1b	You do several trials. Would this many reds come out every time? Why?
Six Trials	Q1c	Six people do trials. Write down how many reds they might get. Why did you choose those numbers?
Range 6	Q2a	In six trials, what might the numbers of reds go from (low to high)?
Range 30	Q2b	In thirty trials, what might the numbers of reds go from (low to high)?
Why on Ranges	Q2c	Why do you think this?
Fifty Trials	Q3a	Out of fifty trials, how many would have 0 Red, 1 Red …10 Red?
	Q3b	Why do you think the numbers you wrote are reasonable?

Table 11

The fourth question addressed the *displaying* aspect, and will be discussed later in this section on the sampling context.

Results for Q1-3: The coding categories for questions 1-3 begins in Table 12 below. Along with a description of the categories, the number and percentage of students (out of the 27 who took the PreSurvey) whose response was coded at a given category are provided. Then some sample responses are given to illustrate selected categories as needed.

After the responses to questions 1-3 are summarized, I'll discuss how the categories from the coding scheme address the framework and help to add detail to the aspects. The same format will be used for reporting on all the PreSurvey results.

Results for Question #1a			
Question Number	Coding Level	Description of Category	No. of Students
Q1ai	2	Either gives a range around 6, such as 5-7, or else writes (for example) "Around 6"	2 (7.4%)
	1	Gives 6 Reds as answer	19 (70.4%)
	0	Gives one number other than 6 Reds, such as 4 Reds	6 (22.2%)
Q1aii	3	Uses proportional reasoning with some explicit statement about what else might happen	2 (7.4%)
	2	Uses proportional reasoning (for example: ratio, average, or percent)	20 (74.1%)
	1	Uses additive reasoning (that is, the number of candies in the jar)	4 (14.8%)
	0	No reason, a vague reason, or an irrelevant reason.	1 (3.7%)

Table 12

The categories assigned to code the results are hierarchical, reflected in the numbers assigned to the coding levels. For example, in Q1ai, the Level 2

responses reflected an appreciation of how results from a single trial might vary around 6 red, whereas the Level 1 responses just provided the expected value with no indication of an expectation of variation. With the Level 0 responses, it's hard to say if someone puts an expected result like "4 Red" because they anticipate variation from the expected value or if they don't even know what is that expected value of 6 Reds. For all tables showing coding levels, a code level of 0 was also assigned if the answer was left blank or if the subject had clearly misunderstood or misread the problem.

For Q1aii, the highest category of responses (Level 3) showed not only that students could correctly reason proportionally, but also that they had a sense of variation around 6 Reds. The lower levels of responses did not include explicit attention to variation. Some representative responses for the top three levels on Q1aii are

[L3] DS: Because 60% are red, so odds are I'd get 6 (Of course I could get more or less)
[L2] BP: Because 3/5 of the candies are red, and 3/5 of 10 is 6
[L1] MM: Because there are more red candies than yellow

Just because a student has a response coded at a lower level does not mean that person lacks the abilities or attributes reflected by the higher levels. The limitations of survey data usually include assessing a response on its own merits. For example, MM's Level 1 response above fails to bring up the proportion of red to yellow candies, and doesn't include attention to variation. Interview data helps look deeper into personal conceptions, whereas the survey data helps me see categories of responses.

Moving next to Q1b, keep in mind that the numbers for the code levels should be thought of primarily as labels, and are not uniform for all questions. In other words, a Level 3 response for Q1aii refers to a certain category which may or may not be different from a Level 3 response in a different question. The focus is on the categories of responses, which is why the tables and examples are provided.

Table 13 below shows the results for Q1b, which was a two-part question. However, both parts were taken together when coding the responses, because experience with the NSF data showed that sometimes a subject would put "Yes" to part one, suggesting that results would be the same every time, but then the same subject in the second part put a response to indicate that what they really meant was results should look similar.

Results for Question #1b

Coding Level	Description of Category	No. of Students
3	Q1bi = No & Q1bii = Explicit reasoning, using both proportional thinking and variation	1 (3.7%)
2	Q1bi = No & Q1bii = Some indication of variation, supported by vague reasoning	16 (59.3%)
1	Q1bi = No & Q1bii = No reason, a vague reason, or an irrelevant reason. Q1bi = Yes & Q1bii = Reasoning explicitly acknowledges variation	9 (33.3%)
0	Q1bi = Yes & Q1bii = No acknowledgement of variation	1 (3.7%)

Table 13

Some representative responses for each of the levels are

[L3] SR: No. With a 60% chance of red and a 40% chance of yellow, your ratios don't change, so it will always vary, but still give red a 20% lead in the jar

[L2] MA: No. There is a likelihood of getting more or less red ones each time.

[L1] GP: No. Because you will probably grab differently and the candies are shifting to different places.
MG: Yes. Because the probability of pulling one kind or another remains the same, with some variation.
[L0] RL: Yes. Returning the candies recreates the original conditions, so the odds don't change.

Although MG says "Yes" to the first part of the question and does mention the stability of the underlying probability of drawing red, he is explicit about expecting "some variation." The mention of variation is what sets MG's response at a higher level from RL's.

Q1c, another two-part question, had subjects pick likely results for six trials on the first part and then offer a reason on the second part. Both parts were taken into consideration for coding purposes, and the categories and results are summarized in Table 14 below. Deciding what would constitute an appropriate choice for the results on six trials involves making a judgment call, and the subcodes attached to this question help identify inappropriate choices as (W)ide, (N)arrow, (H)igh or (L)ow. Only inappropriate choices or blank answers were coded at Level 0.

Results for Question #1c

Coding Level	Description of Category	No. of Students
3	Appropriate choice on Q1ci & Explanation explicitly involves proportional reasoning as well as variation	2 (7.4%)
2	Appropriate choice on Q1ci & Explanation reflects proportional reasoning or notions of spread	13 (48.1%)
1	Appropriate choice on Q1ci & Explanation left blank or lacks any specific reasons relating to details of the distribution	6 (22.2%)
0	Inappropriate choice on Q1ci. W(ide) = Range > 7, N(arrow) = Range < 2, H(igh) = Choices > 5, L(ow) = Choices < 7	6 (22.2%)

Table 14

Examples of responses at the different coding levels are

[L3] RL: {4, 5, 6, 6, 7, 8} Reality does not obey the estimates of probability, so while 6 red candies remains the average outcome, variation is likely.
[L2] SX: {5, 6, 6, 7, 7, 8} All are close to 6 or 6
[L1] JB: {4, 5, 6, 6, 6, 8} The distribution seems about right.
[L0] CS: {4, 4, 4, 5, 5, 5 = (L)ow } Hard to say. Never exact

The table does not break down, for the inappropriate choices at Level 0, how the subcodes were distributed. Two of the Level 0 responses were only (N)arrow - "6, 6, 6, 6, 6, 6" - and two were (L)ow and (N)arrow, such as "4, 4, 4, 5, 5, 5". The other two students left Q1ci blank.

In Q2, the first two parts (Q2a & Q2b) asked subjects to suggest a reasonable minimum and maximum for the results on 6 and 30 trials, and the third part (Q2c) asked subjects for a justification for their choices. Table 15 summarizes the coding procedures and results for Q2a and Q2b:

Results for Question #2a & 2b

Question Number	Coding Level	Description of Category	No. of Students
Q2a	1	Appropriate Choice: Min (3 to 5) – Max (7 to 9)	13 (48.1%)
	0	Inappropriate Choice: W(ide) = Range > 7, N(arrow) = Range <2, H(igh) = Min > 5 or Max > 9, L(ow) = Min < 3 or Max < 7	14 (51.9%)
Q2b	1	Appropriate Choice: Min (2 to 4) – Max (8 to 10)	9 (33.3%)
	0	Inappropriate Choice: W(ide) = Range > 7, N(arrow) = Range < 2, H(igh) = Min > 4, L(ow) = Min < 3 or Max < 8	18 (66.7%)

Table 15

For the subcodes, on Q2a there were five inappropriate choices that were (W)ide, and these choices included "0 – 10", "1 – 10", and "2- 10." Wide choices for Q2a were also automatically (L)ow, but there were six choices that were strictly (L)ow, such as "2 – 8". There were also three responses coded at

Level 0 because one was left blank and the other two showed the subjects had misunderstood or misread the problem. For the purposes of future discussion, any Level 0 responses not specifically accounted for can be assumed to be blank or irrelevant.

On Q2b, there were 9 inappropriate choices subcoded as (W)ide as well as (L)ow. Also on Q2b, one choice was strictly (H)igh, one choice was strictly (L)ow, and the following example was both (L)ow and (N)arrow: "3 – 5".

Reasons for their choices of range were given in Q2c, and Table 16 below summarizes the categories and results.

Results for Question #2c		
Coding Level	Description of Category	Number of Students
3	Explicit mention of increased # of trials leading to increased range, with additional details describing the distribution	6 (22.2%)
2	Explicit mention of increased # of trials leading to increased range	5 (18.5%)
1	Reasoning about one or both ends of the range without explicitly tying the results to the increasing number of trials	4 (14.8%)
0	No answer or irrelevant answer	12 (44.4%)

Table 16

Examples of responses at each of the coding levels are

[L3] CS: The more people used, the more of a variation you will get. Low of 2 to a high of 7 is very different and could happen when using so many people.
[L2] SX: The more groups are pulled, the more chance there is for variation in what the groups consist of.
[L1] SP: Just seems unlikely to pull all reds or all yellows
[L0] BP: The chances of getting any amount of reds is pretty probable

The 11 responses coded at either Level 2 or 3 showed some appreciation for the point of Q2, which was the effect of the number of trials on the range.

The progression in Q1 and Q2 went from one trial to several, then from six to thirty, and finally in Q3 the subjects are asked about results for 50 trials. The first part of the question (Q3a) asked subjects to write down how many of the fifty trials would result in 0 Red, 1 Red, etcetera. I analyzed their frequency charts according to the reasonableness of three characteristics:

(M)ode: Should be at 5, 6, or 7
(E)xtremes: Should not be below 1, and no more than two trials at 10
(D)istribution: Should not be too uniform, or too skewed, or too clustered (i.e. Have a modal frequency higher than 25)

There is subjectivity involved in deciding what characteristics are reasonable, and I made my decisions after considering statistical predictions for 50 trials and also examining multiple simulations of 50 trials using the Fathom software. Table 17 gives the categories and results for Q3a:

Results for Question #3a		
Coding Level	Description of Category	Number of Students
3	All three characteristics are reasonable	2 (7.4%)
2	Exactly two of the three characteristics are reasonable	16 (59.3%)
1	Only one of the three characteristics is reasonable	4 (14.8%)
0	No answer, or none of the three characteristics is reasonable	5 (18.5%)

Table 17

As Table 17 shows, most of the class had at least one unreasonable characteristic. The characteristic that was most often unreasonable was the (E)xtremes, with 14 of the choices for “Pick 50” being too low or having too many low values, and 2 of the choices having too many high values. Also, 6 responses had an unreasonable distribution, and 2 had an unreasonable mode.

In explaining reasons for their choices (Q3b), the categories are similar to those used in earlier questions, and are presented along with results in Table 18 below.

Results for Question #3b

Coding Level	Description of Category	No. of Students
3	Mentions the shape of the distribution, or uses proportional reasoning with some explicit statement about spread	8 (29.6%)
2	Uses proportional reasoning (for example: ratio, average, or percent) or vague reference to spread	6 (22.2%)
1	Uses additive reasoning (that is, the number of candies in the jar)	0
0	No reason, a vague reason, or an irrelevant reason	13 (48.1%)

Table 18

I included the additive reasoning category in the coding scheme for this part of the question because I thought some responses might refer strictly to the numbers of candies in the jar, but none did. Examples of responses from the other categories are:

[L3] RL: A bell curve represents the most likely scenario – the extremes aren't seen often, the average is seen most often
[L2] SX: The most amount of people will get close to 6 because 6/10 of the candies are red
[L0] JB: I am not sure, just looks right

For the people who wrote Level 0 responses about guessing or not knowing why they made their choices, I wonder if they lacked the experience or the language to adequately describe their thinking. Then again, it can be easier in a survey response to put something to the effect of "I don't know" instead of taking the time to think and write about one's reasoning strategies.

Application of Q1-Q3 to Framework (Expecting): The two aspects that Q1-Q3 addressed were *expecting* and *interpreting*. First looking at the *expecting* aspect, I'll comment on the themes that arose for the dimension of *what* was expected, and then discuss themes for the dimension of *why*.

Three themes for responses about *what* was expected came out of my analysis of Q1-Q3, and those themes are listed below and then described:

(i) Concerning Expected Value
(ii) Concerning Repeated Values
(iii) Concerning Range or Extremes

The descriptions offered at this point are tentative, and they continue to develop as the rest of the survey data is summarized. However, it is important to pause and comment on what I saw in the responses so far that helped illuminate the framework.

Many responses showing *what* was expected concerned the expected value (6 Reds for Q1-Q3). For example, most people in Q1ai (One Trial) listed either 6 Reds or gave a range around 6 Reds for what they would expect on one trial. In Q1c (Six Trials), most subjects did give appropriate choices, which included the expected value as well as some higher and lower values. In Q3 (Fifty Trials), for most subjects the mode was either at 6 Reds or close to 6 Reds (5 or 7 Reds). Even in their explanations *why*, subjects often gave clues about *what* they expected, saying for example that results should be close to 6 Red, or that 6 Red would be an average of results. I could see, within the dimension of *what* was expected, that many responses had a common theme concerning the expected value. While not a surprising

theme to emerge, it was instructive to see the responses in Q1 – Q3 highlight the theme.

A second theme, this time concerning repeated values, emerged as I noticed responses that cautioned how results in general would not be the same. Consider these Level 2 responses to Q1b, explaining why results won't be the same each time

SX: Variation – You'll usually pick a different combination
SP: Each session could produce different results
SC: It can't possibly always be the same.

The responses themselves talk about how results would be different, but I have phrased the theme in terms of the how likely or unlikely results would be to repeat. Clearly some repetition of results is expected in 6 or 50 trials, since there were multiple outcomes of 6 Reds, for example. For instance, some students put all distinct values for their 6 trials (such as "3, 4, 5, 6, 7, 8"), but most choices did have some repeated values (such as "3, 5, 6, 6, 6, 8"). The theme I noticed concerning repeated values suggests that responses do give a sense for how much or little repetition is expected in results.

The third theme has to do with the range, or extreme values. That is, some responses clearly indicate *what* is expected in terms of a range, or in terms of extreme results, such as

(Q1b) JL: One pull may give you 100% reds, the next 10% reds, and so on
(Q2c) EM: I figure at some point one of the 30 has to draw only 1 red and at some point one of the 30 has to draw all 10 of red
(Q3c) SP: Seems that the majority of people would get at least 2 but no more than 8 reds

I noticed that for the range predictions on Q2, many people on Q2b ("Range 30") put inappropriate choices which were wide and low, going from 0 to 10 reds. Such an expectation shows an overanticipation of extreme results for the sampling situation of Q1-Q3, because a set of thirty trials is not likely to show results below 3 Reds.

Three themes for responses about *why* students held their expectations also came out of my analysis of Q1-Q3, and those themes are listed below and then described:

(i) Involves Possibilities or Likelihoods
(ii) Involves Variation or Distribution
(iii) Involves Proportional Reasoning

For the theme involving possibilities or likelihoods, the emphasis I focused on was the general language with which subjects used to justify why they expected certain results. Here are some sample responses that helped me identify the theme involving possibilities or likelihoods:

(Q1aii) CM: It's possible to get 10 yellow
(Q1aii) DM: The likelihood of the ratio being 6:4 is high
(Q1bii) JL: The likelihood that every grab yields 60% reds is just not there.
(Q2c) SR: It is not likely, but not impossible to reach in and get all red candies, but there is a better chance not to

Other language beyond what was possible or impossible, likely or unlikely included the subjective notion of greater of lesser chances of obtaining certain results.

A second theme for reasons *why* subjects held their expectations had to do with variation or the distribution. For example, in Q1b when explaining

why results would be different, some students wrote their in explanations how the results should vary:

BP: Because it is random, and the number can vary
JX : It would vary because the candy is not in any order
LW: The number could vary

For many responses with a theme involving variation or distribution as a part of students' explanations, the *what* and the *why* are somewhat intertwined. For example, consider DS's explanation in Q3b about *why* she had put her choices for the fifty trials in Q3a:

(Q3b) DS: Because most people would be close to the 60% of the total # of reds. Fewer people would be at the far ends of the curve (a lot higher or a lot lower than 60%)

DS wrote the above response to justify her earlier choices, and her explanation exhibits some sound facets of distributional reasoning. We can see in her response *what* she expects, but the emphasis or purpose of her statement concerns *why*.

While the other two themes at this point are still emergent, and dependent on future responses in the survey data to add richness, the third theme (proportional reasoning) came through strongly and clearly as a part of the responses for *why.* Here are some examples:

(Q1aii) TO: Because the ratio of red to yellow is 6 to 4
(Q1aii) SR: Since you have 60 red and 40 yellow, you have a 60% chance that each candy is red
(Q1c) DM: Because they are mixed with a 60:40 ratio
(Q1c) AL: It's a 6:10 chance you will pick a red

Proportional reasoning was a feature of many categories in the previous

tables, and for many subjects the underlying proportion of candies was clearly a key component of their reasoning *why* they held their expectations.

Application of Q1-Q3 to Framework (Interpreting): There were different themes that came out of my analysis for Q1-Q3 having to do with the dimensions of *causes* and *effects* of variation, as well as the dimension of *influencing expectations and variation*.

Looking at *causes* of variation, one theme had to do with the physical environment of the sampling situation. Two ways in which the theme emerged were the way the candies were mixed and the way the handfuls were pulled. For example, consider these responses on Q1b:

DM: Because on occasion you will have varying answers based on how many red or yellow happen to be in the area you grab
EM: Because yellow candies could be bunched together in the jar & our handful could have a lot more yellow or a lot more reds
GP: Because you will probably grab differently and the candies are shifting to different places

Other responses talked about how the candies were “mixed” or “jumbled.”

Considering *effects* of variation, one theme was how reality does not always match with what probability suggests. For example, in Q1aii (“One Trial”), 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 include:

(Q1b) DS: Because probably outcomes aren't for sure outcomes
(Q1c) RL: Reality does not obey the estimates of probability
(Q1c) SR: 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. In identifying this idea as a theme having to do with effects of variation, I do not suggest that the subjects necessarily see it as such. That is, they may or may not identify the tension between prediction and outcome as an effect of variation. Rather, I am classifying this theme as an effect because of how I see the situation: Namely, the sampling situation inherently involves variation, which leads to unpredictable individual results. If there were no variation, then the subjects' predictions would always match the actual outcome. Since there is variation, an effect is that reality does not always match the prediction.

The reasoning I just described about an *effect* of variation leads to a second theme, which is the way that variation impedes making a prediction. That is, some students suggested that it was hard to say what would result, as these sample responses show:

(Q1b) AL: You can make a prediction, but not a concrete answer as to what color you will pick
(Q1b) LT: Always getting six red candies is hard to predict
(Q1c) CS: Hard to say. The odds are never exact

Also, many responses mentioned guessing in the sense that they couldn't

know for sure, and therefore a guess was the best they could offer. Again, I the above ideas as a theme for the *effect* of variation. It is true that uncertainty goes alongside the themes I've just discussed for the *effects* of variation, and one could just as easily point to difficulties in making a prediction as an effect of randomness or uncertainty as one could point to an effect of variation. However, semantics do come into play when talking about uncertainty, randomness, and variation, and I am referring to the above themes as an *effect* given the broad definition of variation that I introduced in first chapter.

Finally, considering the dimension of *influencing expectations and variation,* two themes that came out concerned the numbers of candies in the jar and also the number of trials performed. The first of these two themes is more tentative, meaning it wasn't clear to me exactly what the relationship might be in the subjects' minds between the numbers of candies and the resultant variation. Here are some sample responses that drew this theme together for me:

(Q1aii) JB: Over half the candy in the jar are red, so the probability that more than half [in the handful] will be red is good
(Q1aii) MA: I stand a greater chance of pulling more red than yellow, because there are more of them to begin with in the [jar]
(Q1aii) RF: Because if I had more red I have more probability to get more of these

The responses shared above all fell into the category of additive thinking, meaning that the subjects only appealed to the quantity of candies in the jar instead of the ratio. Also, the subjects offered these additive responses as

an explanation for *why* they put their previous expectations. The reason I've included the theme of additive reasoning (which I am just calling the numbers of candies in the jar) as a part of the dimension of *influencing expectation and variation* is because I wonder if the subjects are thinking that if the jar had even more candies (but a consistent ratio), then would the probability of choosing red also increase from the smaller jar to the larger jar?

The second theme concerning the number of trials was more robust, since all of Q2 was essentially aimed at the connection between increasing trials and increasing range. As Table 9 showed, 11 students wrote responses clearly citing this connection (at Levels 2 and 3) , and here are some additional responses:

(Q2c) CM: The range will increase with increasing attempts
(Q2c) JL: The more people that do the experiment, the more varied the results.
(Q2c) RL: As the number of trials goes up, so expands the range of possible outcomes towards the extremes.

However, even though the above responses show a generally correct thinking about the influence of the number of trials on variation, there is still a question about how readily the range would expand. For instance, on Q2c (Range 30), many people put "0-10", the maximum range possible, when in fact the lowest results of 0 and 1 are very unlikely to occur in a set of thirty trials. Still, the general idea that more trials might expand the range is an important theme in the dimension of *influencing expectation and variation.*

Questions 1-3 gave me my first look at how responses might fit into

my framework. As future questions also informed the aspects of *expecting* and *interpreting*, I'll mention details that continued to add depth to these two aspects. The next question, Q4, had more to do with the *displaying* aspect, and spanned the contexts of sampling and graphs. I've included the question in the section reporting on the sampling context mainly because the question continued using the same trial from Q1-3: Handfuls of ten drawn from the jar containing 60 red and 40 yellow candies. However, Q4 asked students to make a graph of what they thought the results from fifty trials might look like. Although Q4 (named "Make Graph") is definitely an extension of Q3 ("Fifty Trials"), where the students filled in a chart showing predicted frequencies, I deliberately put Q4 on a separate page to see if students would really go back and just use the information on their charts to construct a graph or if they would just make a graph independent of their previous predictions.

Results for Q4: Of the 26 completed graphs, 14 clearly matched the chart of Q3, and 2 clearly did not. On the other 10 graphs, it was difficult to tell whether or not they were drawn with the intention of reflecting Q3's numbers. An example of this difficulty is shown by GP's graph (see Figure 16 on the next page). GP had put on Q3, for fifty trials, these frequencies for 0 Red, 1 Red, on up to 10 Red: 1, 2, 4, 5, 6, 7, 10, 7, 5, 2, 1. His reasoning on Q3 was that the "top of the pyramid is 6 or the most probable and it just cascades down." Although his choices go too far into the low extremes, his description of the shape of the distribution is fairly reasonable and does conform to his graph on Q4. But without the vertical scale provided, it's hard to say for sure

if in fact his graph reflects his frequency chart.

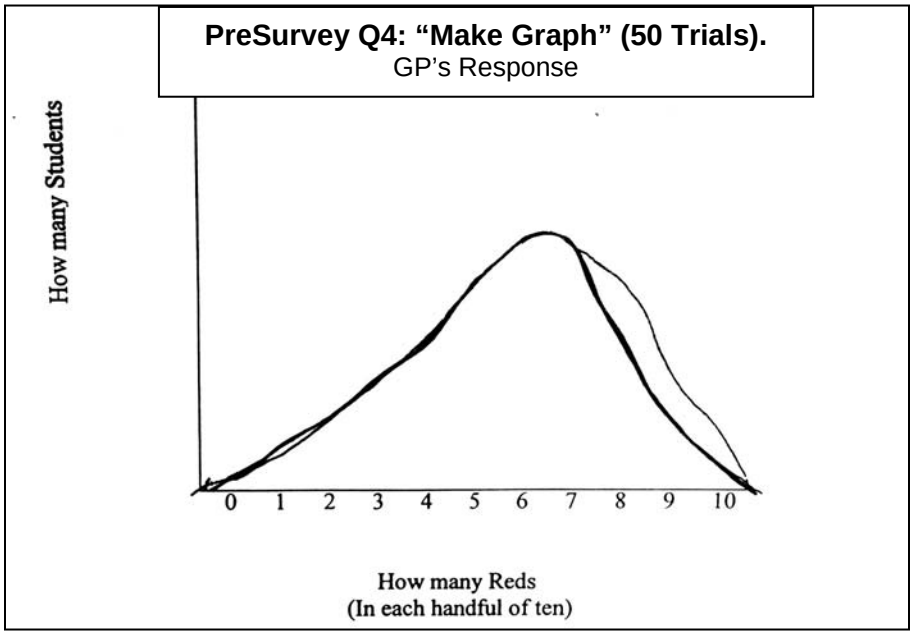


Figure 16

An example of a graph that clearly matched the student's chart is given by SP in Figure 17 (her chart had frequencies of 0, 1, 5, 5, 10, 10, 8, 8, 1, 0):

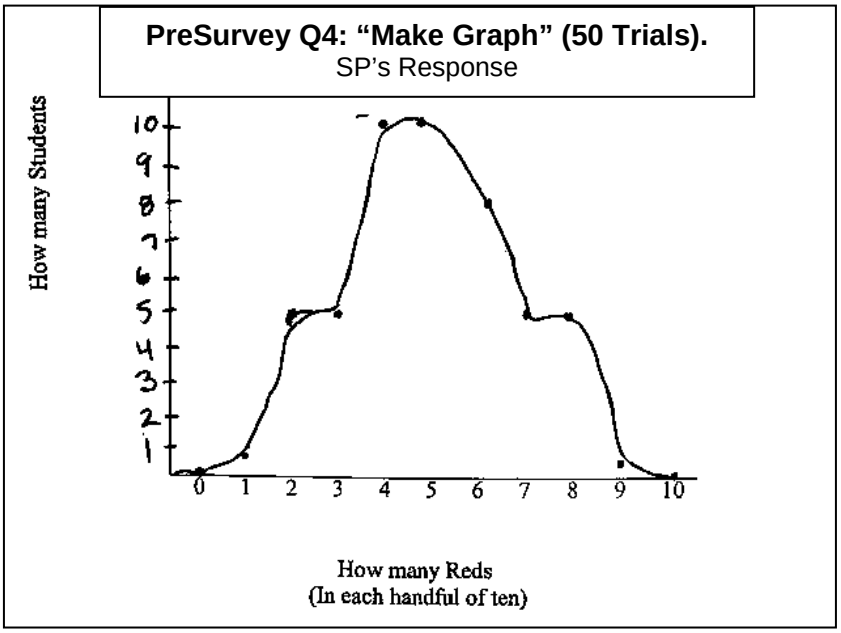


Figure 17

SP had placed distinct points in her graph corresponding to her choices on the frequency chart of Q3. A graph that clearly did not match the prediction was given by MG. MG, whose frequency list on Q3 was reasonable, gave a sort of a scatterplot which did not make much sense except for the showing that MG was indeed trying to account for all the fifty trials.

In addition to seeing if their graphs matched their charts, the responses on Q4 were coded according to the categories listed below in Table 18:

Results for Question #4

Coding Level	Description of Category	Number of Students
3	Reasonable variation around the center, vertical scale is also good.	1 (3.7%)
2	Approximate shape for the distribution, centered at or close to 6, but having too much or too little variation. Possibly no attention to vertical scale	17 (63.0%)
1	Attending to every trial (as in a scatterplot or ordinal graph), or graphs in the wrong place w/ no attention to vertical scale	6 (22.2%)
0	No answer or an attempt which is extremely difficult to decipher	3 (11.1%)

Table 18

MG's graph (not pictured) was a scatterplot coded at Level 1 because even though it was hard to decipher, the graph showed an attempt to account for every trial. Most responses were coded at Level 2, as were GP's and SP's graphs shared earlier. GP had too much variation and no vertical scale, but a good skewed-bell shape. SP's graph was a little off-center, but had a plausible shape and a detailed vertical scale. An example of a Level 3 graph was given by SW in Figure 18 on the next page. Although she did have both a high and

low extreme, in every other respect her graph was quite reasonable, and the added detail of the vertical scale made SW's graph stand out from the rest. SW's graph also clearly reflected her choices made in Q3 on the frequency chart.

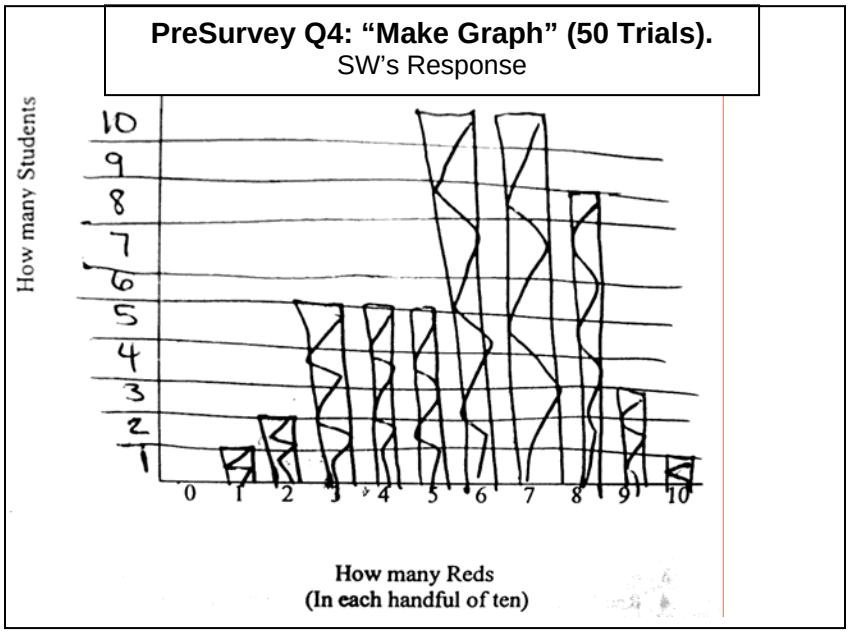


Figure 18

Graph sense plays a major part in Q4, and the lack of familiarity with the need for detailed axes came through in the Level 2 graphs which had a reasonable shape but lacked any sort of scale to tell how many trials were occurring at each possible result.

Application of Q4 to Framework (Displaying): Q4 appealed to the dimension of *producing graphs*, and the two themes that came out of the responses to Q4 were reinforced by the two other questions on subsequent surveys for which students came up with their own graphs. Those two themes

have to do with the technical details (type of graph used and appropriateness of scales along the axes) and with the characteristics of the distribution (the reasonableness of the center, spread, and shape of the graph).

Concerning the technical details, most students chose a smooth curve with a reasonable shape, but there were other types of graphs: Bar graphs, lineplots, and scatterplots. As for the vertical axis, some students added sufficient detail while others ignored the axis completely.

Regarding the characteristics of the distribution, most students had reasonable modes, with some going too low and just one going too high. The spread were overwhelmingly wide, because the graphs included or seemed to include the highly unlikely 0 red outcome. A few graphs were too narrow, and a few showed reasonable spread. Finally, in terms of overall shape, most were reasonable, albeit showing more symmetry than might actually occur.

What I learned from Q4 was that the notion of a bell-shaped curve, or other symmetric shape (such as the inverted “V” shape that a couple of students drew) was a dominant part of the classwide thinking. The thinking actually comes out first in the explanations for Q3, where subjects wrote about the higher frequencies being near 6 Reds, and the lower frequencies moving towards the ends of the range. However, the type of graph chosen and the attention to scales really affect the end product of coming up with a graph to adequately portray what might happen in fifty trials. Moreover, just as the frequency charts for Q3 proved to be mostly wide, so too were most of the graphs in Q4 wide. For the purposes of my framework, I was curious to see

what I could pay attention to in general as I considered how subjects *produced* graphs to display the variation they expected. I found the themes presented above to be a way to look at and categorize not just the graphs in Q4, but also the graphs on the later surveys.

Data and Graphs

Although Q4 involved *producing* graphs in the aspect of *displaying* variation, I included Q4 in the reporting of the other questions that were based on the sampling context. The next two PreSurvey questions, Q5 – Q6, focused on the context of data and graph, and also appealed to the *displaying* aspect. However, the dimensions that responses for Q5 and Q6 related to were *evaluating and comparing graphs* and also *making conclusions about graphs*. Again, I'll share the questions, categories of responses, and then discuss the application to the framework.

Results for Q5 & Q6: In Q5a, two graphs were shown for the test scores of two different classes (the Yellow Class and the Brown Class) which had taken the same version of a test. A similar situation followed for Q5b with the Pink Class and the Black class. In Q6, two graphs showed the heights of students at two different school (School A and School B).

PreSurvey Questions #5 & 6		
Name of Task	Question Number	Description of Task
Classes Y & B	Q5a	Did the two classes do equally well, or did one class do better?
Classes P & B	Q5b	Did the two classes do equally well, or did one class do better?
Schools A & B	Q6	Which graph shows more variability in students' heights?

Table 20

The text of the questions are listed in Table 20 above, and the graphs which accompanied the questions are presented along with the discussion of the results for each question or subquestion.

Results for Q5 & Q6: Since Q5a is the first to be summarized, the graphs used for the Yellow and Brown Classes are given in Figure 19:

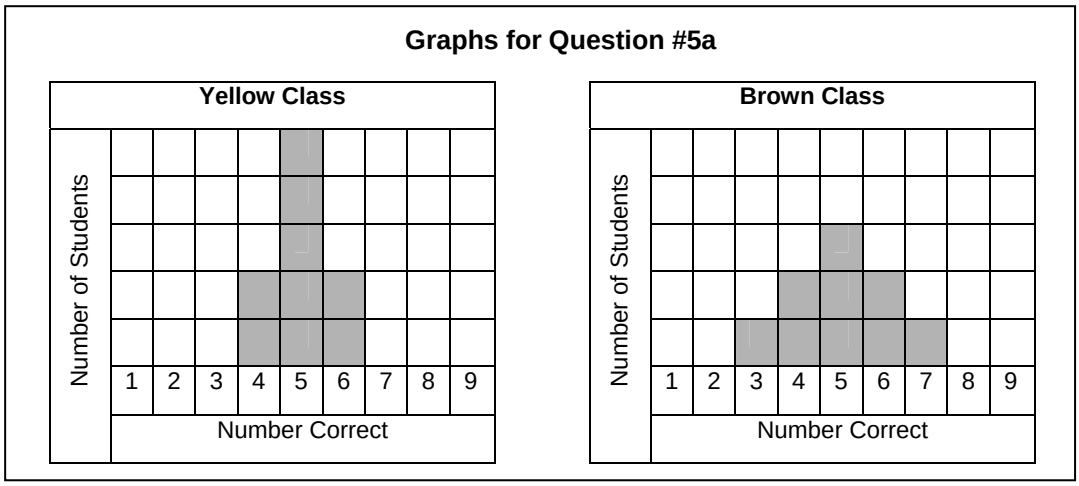


Figure 19

Note that the total number of students for each class was the same, as was the total class score (45 correct), and the class mean, median, and mode (5 correct). The coding categories and numbers of responses start in Table 21. Most of the students did use the mean as a way of comparing the two classes, although two of the subjects had calculated the average incorrectly. Some sample responses from the top four categories include:

[L4] SC: The two classes taken as 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 seeing 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.

[L3] JM: In terms of raw numbers, the two classes got an equal number correct. The Yellow class had a distribution that fewer students getting lower and higher correct answers. The Brown class had a more even distribution

[L2] CM: The average score for each class was 5. They did equally well.

[L1] DM: They both did equally as well, because each class got 45 correct.

Results for Question #5a

Coding Level	Description of Category	Number of Students
4	Sophisticated use of Average and Variation	2 (7.4%)
3	Variation explanation, spread or tightness of distributions	4 (14.8%)
2	Using average (even if the actual calculation for the average was incorrect)	12 (44.4%)
1	Using the sum of the class scores (even if the actual calculation for the sum was incorrect), or individual graph characteristics.	7 (25.9%)
0	No answer or unclear reasoning	2 (7.4%)

Table 21

The above responses show the progression up through the categories quite well: DM (L1) relied on the sum total of the class scores, while CM (L2) used only the average score. JM (L3) appealed to the distribution, and SC (L4) used both average and spread in her explanation. The symmetry of both distributions no doubt made it easier for SC to see the "restacking" she wrote about, and a similar idea was actually shown on the paper by GP (see Figure 20 on the next page).

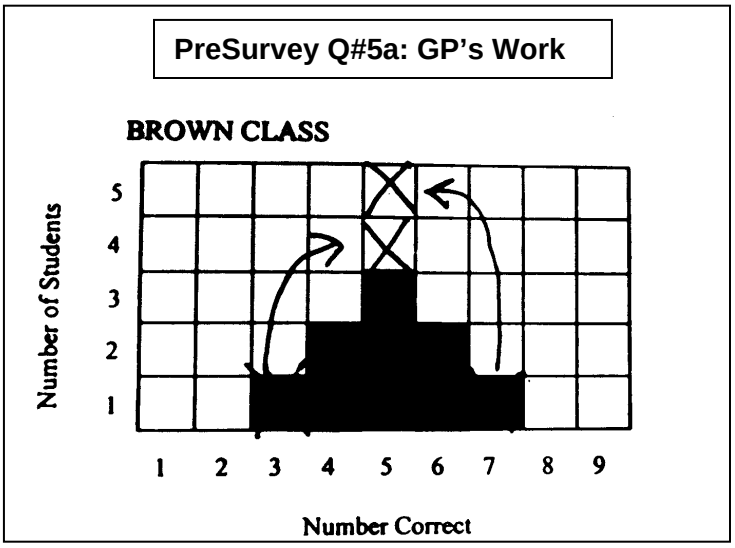


Figure 20

The class sizes were different in Q5b, and one of the graphs (see Figure 21) was not symmetrical. However, regardless of the type of measure used, the only justifiable conclusion about the Pink and Black classes was that the Black class did better.

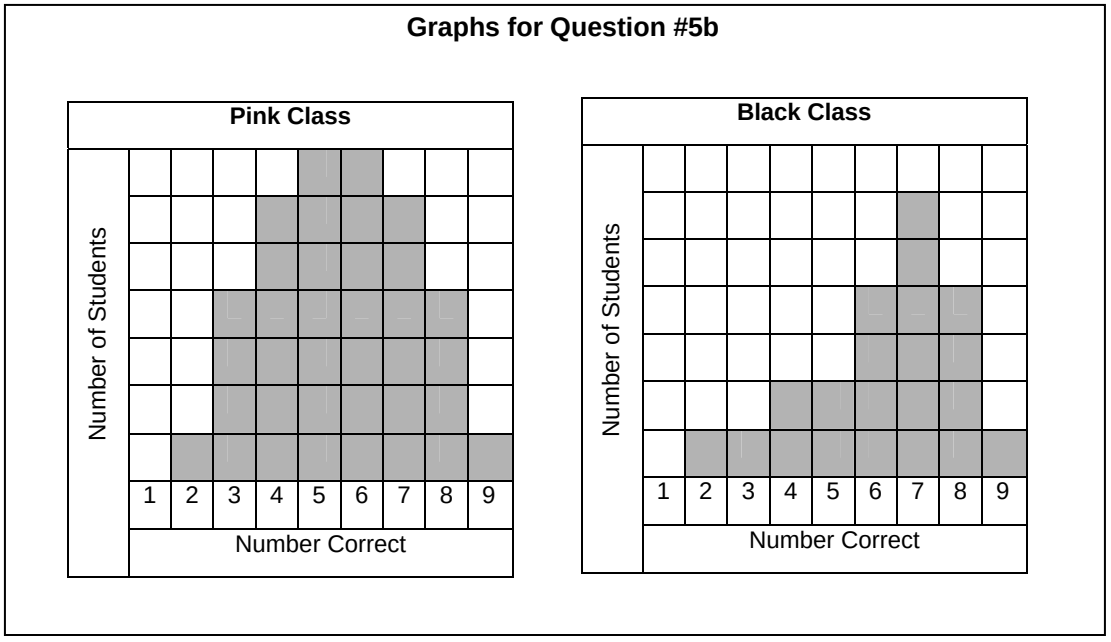


Figure 21

The mean for the Pink Class is 5.5 correct (198 total correct / 36 total students), while the mean for the Black Class is 6.2 correct (130 total correct / 21 total students). Pink's median is 5.5 correct, and Black's median is 7 correct. In asking this question, it was hoped that students would not solely rely on the average, but also consider the distribution of the scores. Therefore, the higher category of responses focused on distributional thinking that showed how the subjects attended to more than just the average. Table 22 summarizes the categories as well as the number of responses at each level.

Results for Question #5b		
Coding Level	Description of Category	Number of Students
3	[Black did Better] Use of characteristics of the distribution besides just the average	4 (14.8%)
2	[Black did Better] Using average only (even if the actual calculation for the average was incorrect)	9 (33.3%)
1	[Black did Better] Using the sum of the class scores, or individual graph characteristics (such as the height of the bars)	0
0	[Pink did Better] or [Both are Equal] or [Black did Better] – If they put the latter, it was supported by weak or unclear reasoning	14 (51.9%)

Table 22

More students either identified the Pink Class as doing better overall or else suggested the Black Class did better but offered a very weak reason.

No students had responses coded at Level 1, meaning that if they did put Black as the better class, they either appealed to the overall distribution (Level 3), or used the average (Level 2) or gave a poor or no reason (Level 0).

Examples of responses at Levels 3, 2, and 0 are:

> [L3] JB: Black class did better with about 75% of the class over the median, while the Pink class has only about 50% of their class over the median

[L2] MG: Black class did better because their average was higher (6.19) opposed to a 5.5 average for the Pink class
[L0] GP: The Black class did better since they had more 6's unchecked by the 5's. The Pink class 6's are checked by their 5's.
CS: The Pink class answered 36 questions correctly and the Black class answered 21 questions correctly
JX: This cannot be determined since the classes do not have the same number of students.

I included several Level 0 responses because they each show different ways of thinking. GP seems to be considering the heights of the bars at 5 and 6 correct, but using the information to questionable effect. CS seems to be confounded by graph sense, in that she has obtained the total of squares in each graph but has not appreciated what that total actually gives. For JX, whereas the Yellow / Brown classes of Q5a could be compared because of having equal numbers of students, the Pink / Black classes on Q5b cannot even be compared because they had different numbers of students.

Q6 also showed graphs containing different numbers of students. Just as Q5a was a relatively easy calculation to make to get the totals and averages but Q5b was more involved, so too was Q6 even more time-consuming to obtain total numbers of students. However, Q5b and Q6 were not designed with the intention of making or even inviting students to do calculations: The highest categories of response for Q5b and Q6 included a correct answer with a justification dependent on some form of reasoning about range, spread, or the overall distribution. The graphs for Q6 are shown in Figure 22, with the heights of students given in centimeters:

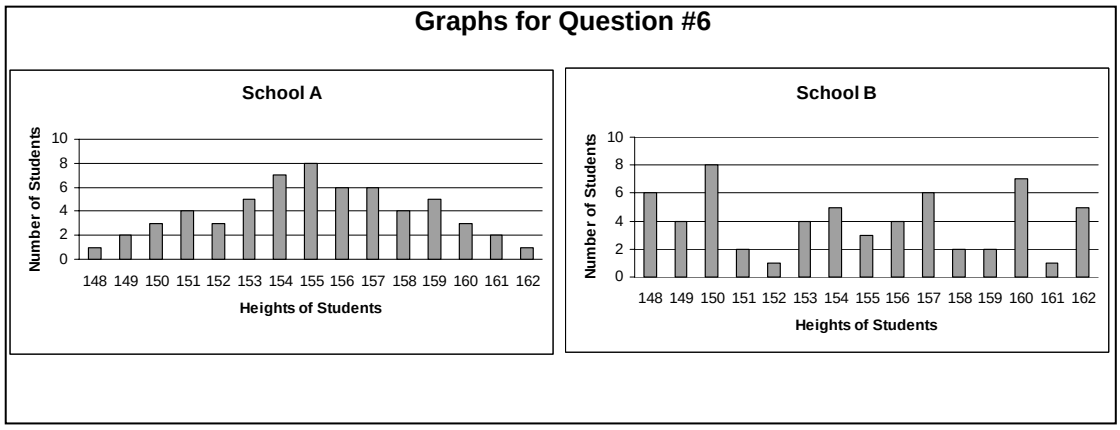


Figure 22

What especially sets Q6 apart from the other questions in the context of data and graphs is that variation is explicitly addressed as something for the students to look for, in the form of the phrasing: "Which graph shows more variability in students' heights?" Table 23 below shows the categories and numbers of responses for Q6.

Results for Question #6		
Coding Level	Description of Category	Number of Students
2	[School A] with a basically correct reason for comparing the schools	14 (51.9%)
1	[School A] with vague or weak reasoning [School B] with some kind of reasonable reason	8 (29.6%)
0	[[School A or B] with no reason, or with incorrect or irrelevant ideas	5 (18.5%)

Table 23

Some examples of responses in each category will show the kinds of reasoning offered by subjects:

[L2] JB: [School A] A has more variability in heights of students, ranging from 145-165. School B ranges from 148 to 162"

[L1] AL: [School A] The heights vary more in A. There are more differences in heights

BP: [School B] There is more variation from the average at School B than at School A.

[L0] RL: [School B] School A may be more homogeneous with regard to ethnicity, which is a big factor in determining height

Most of the Level 2 reasons given for choosing School A pointed to the wider range in comparison to the more narrow range of School B, as exemplified by JB's response above. A possibly confounding issue may have directed some the students to an erroneous conclusion, and the issue is over the context and wording of the task. This specific task has been used elsewhere besides the NSF grant (Torok & Watson, 2000), and I too was initially thinking in a different direction than asked for when I first read the task. School B seems to show a greater bar-to-bar fluctuation of heights. That is, the heights of the adjacent bars vary up and down more frequently than the smooth rise of the bars in School A. Visually, the heights of the bars can tend to attract one's focus. Also, the language and context of the problem addresses heights of the students at the different schools. Thus, there are heights of the students (context of the problem) and also heights of the bars (visual presentation of the data) to consider, potentially causing some confusion as JX shows:

[L1] JX: [School B] Because there is a wider range of difference in the heights recorded in B than A. B's lowest height is the same as A, while B also has 4 groups that are taller than the tallest group in A

When JX talks of "B's lowest height", this is not in reference to a student height but to an actual height of a bar. JX exemplifies what I meant earlier about distinguishing between these two uses of "height." A careful consideration of what the task is addressing seems required, as well as a good sense of graphicacy, when discussing variation in the contexts of data and graphs.

Application of Q5 & Q6 to Framework (Displaying): The two dimensions that Q5 and Q6 related to in the *displaying* aspect were *evaluating and comparing graphs* and also *making conclusions about graphs*. These dimensions, along with the themes within them that I noticed coming from the responses to Q5 and Q6, are next discussed.

In *evaluating and comparing graphs*, the themes I noticed were that responses had any of three different foci:

(i) Focus on average
(ii) Focus on spread
(iii) Focus on shape or distribution

For example, in Q5a and Q5b, many responses used the average, and not only the mean was mentioned: Some students also used the mode or median. Only one student in Q6 mentioned the average, and it's doubtful that value was actually calculated because it's a lengthy computation. Consideration of the average is appropriate for Q5 and Q6, but I was also interested in responses that took into account some measure of spread. Most of the Level 2 responses in Q6 did have a range focus, as these additional sample responses show:

GP: [School A] Since it goes from 145 to 165. School B goes from 148 to 162
DS: [School A] Because the heights vary all way from 145-165 cms but only 148-162 cms in School B (not as broad a variation of heights)

The shape of the graph or the way the data was distributed was reflected in some of the responses for Q5 and Q6. In Q5a, SC and GP's ideas were

highlighted to show how data could be redistributed on one of the graphs to obtain the other graph, and JM's response showed attention to the distribution of both the Yellow and Brown class. JB's Level 3 response on Q5b showed good reasoning about the distribution of data for the Pink and Black classes, as did the following example:

> MM: [Black Class] The Pink class had more students scoring in the mid level range, whereas the Black class had a FEW students scoring below mid range and MOST of the students scoring in the mid – to – high range of # correct.

In Q6 there were also a few responses that considered the distribution of the data within the range, as SC shows:

> SC: [School A] At School B there were no students in the extreme low or high range in heights. They tended to "cluster." Whereas School B students had more variability with some students measuring into each category.

There weren't many responses to Q5 or Q6 that specifically used language to describe the shapes of the graphs, but I thought that the questions provided a good opportunity for subjects to talk about shape. Earlier, in Q3 when explaining their choices for fifty trials in the sampling context, there were some references to "bell curve" and "top of the pyramid," and so I had expected similar descriptive language for Q5 and Q6.

For the dimension of *making conclusions about graphs*, there were two tentative themes that began to emerge. One theme emphasized the level of detail and usefulness of the graphs, and the other theme emphasized making decisions in context. They were both tentative themes at this point in the analysis, because very few responses addressed them. However, the aim in

thinking about themes for the dimensions of the framework was not only to draw quantitative support, but also to think about relevant issues that I could consider in analyzing future responses.

What got me thinking about the level of detail and usefulness of the graphs was the notion of graph sense and how that sense plays such a pivotal role in questions involving data and graphs. After all, to discuss variation in a graph, one needs to make sense of the graph, and some graphs are more readable and useable than others. On Q5a, one subject wrote “I don’t understand how the graph is laid out” and went on to say “the graph doesn’t make sense to me.” Similarly, the details of what the graphs actually meant escaped some students on Q5b, especially for those who noted that the Pink class did better:

LW: [Pink Class] The Pink class had more correct words spelled as a class than the Black class
SC: [Pink Class] Just squint. You can see that the Pink class had many more correct answers than the Black class
SR: [Pink Class] Pink class did much better. First I could visually see it, then I added it all together.

As mentioned previously, there is virtually no plausible way that a correct reading of the two graphs could support a conclusion that the Pink class did better than the Black class. Graphs cannot be useful if they are not understood, and I wonder about the level of detail and the graph sense required to make supportable conclusions when dealing with data and graphs. I already drew attention to a possibly confounding issue in Q6, again involving the details and usefulness of the graphs.

There were not many responses that addressed the other theme of emphasizing decisions in context, but the few that did came from Q5a and Q6. DS noted in part of her response to Q5a: "Yellow does have more people 5 or higher which as a teacher I think is better. You know the whole class is getting what you're teaching." She did also use averages to compare the classes in her answer, but the above quote shows how she was attending to the hypothetical situation from which the data was supposed to have come. Similarly, when RL talked in his Q6 response about the ethnicity of the students being related to height, he was trying to think realistically about where the data came from. JL also attended to contextual factors in her response to Q6, noting that "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."

To summarize, what I gained from Q5 and Q6 for the purposes of thinking about the framework all related to the *displaying* aspect. In *evaluating and comparing graphs*, it does seem that something useful to pay attention to in looking future responses is how the responses run along three themes: The focus on average, the focus on spread, and the focus on shape or distribution. Regarding *making conclusions about graphs*, two themes to be mindful of in future responses is how the responses emphasize the level of detail and usefulness of the graphs, and also how the responses emphasize decisions made in context.

Probability

The last three questions on the PreSurvey all were in the probability context. Q7 helped address the aspects of *expecting* and *interpreting*, and was crafted to resemble Q1 in structure. Q7 will be discussed next, and separately from Q8 and Q9, because Q8 and Q9 were very different from Q7. A single trial for Q7 consisted of fifty flips of a fair coin, noting how often the coin landed heads-up, and the expected value for Q7 is 25 heads.

The questions asked in Q7 are presented in Table 24 on the next page. The subquestions are almost identical in emphasis to those asked in Q1 for the sampling context. One difference is in Q7b, where the question simply asked how the second trial would compare to the first trial (in Q1b, several trials were hypothesized). Where I had named Q1b "Several Trials", I named Q7b "Compare Trials".

PreSurvey Question #7		
Name of Task	Question Number	Description of Task
One Trial	Q7ai	Mark does one trial. How many heads do you think he might get?
	Q7aii	Why do you think this?
Compare Trials	Q7b	Mark does a second trial. (Q7bi) How do his results on the second trial compare with the results on the first trial? (Q7bii) Why?
Six Trials	Q1c	Mark then does six trials. Write down how many heads you think he might get on each trial. Why did you choose those numbers?

Table 24

One reason for offering essentially parallel questions but in different contexts was to provide additional opportunities for students to put down what they thought. It could be that a student may not write much in a sampling context but may write quite a bit more in a probability context (or vice versa). That is,

people can relate differently to the two contexts. A related reason for the isomorphism was to assess any classwide shifts in responses across contexts, and this assessment takes place after all the survey data has been summarized.

Results for Q7: The categories and numbers of responses start with Table 25. The categories for Q7a were similar to those used in Q1a. One difference was in a category from Q1aii to Q7aii (Level 1): I had thought some responses might offer a form of additive reasoning for the coin flipping, perhaps by focusing on the two equal sides of the coin. Instead, there were some Level 1 responses that weren't quite additive, nor did they invoke proportional reasoning, but they seemed potentially reasonable even though they lacked specificity.

Results for Question #7a			
Question Number	Coding Level	Description of Category	No. of Students
Q7ai	2	Either gives a range around 25 such as 22-28, or else writes (for example) "Around 25"	1 (3.7%)
	1	Gives 25 heads as answer	21 (77.8%)
	0	Gives one number other than 25 heads, such as 23 heads	5 (18.5%)
Q7aii	3	Uses proportional reasoning with some explicit statement about what else might happen	2 (7.4%)
	2	Uses proportional reasoning (for example: ratio, average, or percent)	17 (63.0%)
	1	Uses additive reasoning , or gives a reasonable response which makes sense but lacks specificity	4 (14.8%)
	0	No reason, a vague reason which makes no sense, or an irrelevant reason.	4 (14.8%)

Table 25

The descriptions for the categories in Q7ai give the idea of the kinds of responses coded at each level, and some examples for the categories in Q7aii

are given below:

[L3] JM: The coin has a 1:2 chance of landing on heads. The more often you flip, the chances of the 1:2 ratio will be closer to that – 1 in 2
[L2] EM: He has a 50% chance of landing on heads
[L1] RF: Because you have the same chances
[L0] SW: Maybe a little more than half 'cause it started on heads: I have no idea really

JM's response uses the ratio defined by the fair coin, but his response also suggests that the cumulative average of many flips approaches that ratio: He shows thinking that aligns with the Law of Large Numbers. EM uses only the proportion in her reasoning, and RF's response sounds reasonable but needs more specific details. SW's answer suggests a possible physical cause of variation.

The structure of the categories in looking at Q7b was different from those categories for Q1b because the two questions were worded differently. In Q1b, students were asked to imagine doing several trials, whereas in Q7b the wording suggested that "…a second set of 50 flips" was done. Students were then asked in Q7b how the results on the second set (or trial) compared with the results on the first set (or trial). The essential theme that the two questions (Q1b and Q7b) get at is the same: Do they expect the same results or not, and why? However, one reason the wording was changed is because this PreSurvey instrument was given in class prior to any instruction, and I wanted the situation to be as free from confusion as possible. Past experience with probability questions where a trial consists of multiple actions (i.e. "Flip a

coin 50 times" or "Spin the spinner 50 times") suggests that once the situation advances to multiple trials – say, 30 trials – some students get confused and think they are now doing 30 flips or spins instead of repeated trials of 50. Drawing several handfuls of ten candies, or six handfuls, or fifty handfuls, may be easier to keep straight in students' minds as opposed to doing thirty sets of fifty flips. Also, the wording on Q1b specifically addresses whether the results would be the same "every time", but on Q7b it asks more generally how the results "will compare." In asking for Q7b "How do you think his results…will compare…?" , I made the decision to interpret responses about how results on the second set might be "similar" as implying "similar but not the same." The categories and numbers of responses for Q7b are given in Table 26, followed by some representative responses from each category.

Results for Question #7b

Coding Level	Description of Category	Number of Students
3	[Different or Similar] w/ Explicit mention of a range or spread	3 (11.1%)
2	[Different or Similar] w/ Some additional information, such as use of ratio, average, percent, or giving specific alternatives for results	11 (40.7%)
1	[Different or Similar] w/ No additional information provided	7 (25.9%)
0	No answer, or emphasizes guessing, not knowing, or how results will be the same	6 (22.2%)

Table 26

As mentioned above, a main point of the question was on whether or not results were predicted to be the same, and that is why a response indicating results would be "different" shared a commonality with responses

about results being “similar” : My assumption was that both kinds of responses implied results might not be the same on both trials, and this assumption also extended to responses of how results would “nearly be the same”, as did MA’s in the first of the sample responses that follow:

[L3] MA: It will be nearly the same, or the same. The variation may be only 2-3 one way or the other
[L2] LT: I think it might come pretty close to it. But, I think one side might be higher
[L1] SX: They will be similar but not the same
[L0] MG: Same. Probability will remain the same.

SX’s response lends credence to my assumption that “similar” connotes “not the same”, and judging by that assumption, most of the class held the idea that results on the second trial would likely not be identical to the first trial.

Q7c was also parallel to Q1c, a two-part question that had subjects pick likely results for six trials on the first part (Q7ci) and then offer a reason on the second part (Q7cii). Both parts were taken into consideration for coding purposes, and the categories and results are summarized in Table 27 below.

Results for Question #7c

Coding Level	Description of Category	No. of Students
3	Appropriate choice on Q7ci & Explanation explicitly involves proportional reasoning as well as variation	2 (7.4%)
2	Appropriate choice on Q7ci & Explanation reflects proportional reasoning or notions of spread	13 (48.1%)
1	Appropriate choice on Q7ci & Explanation left blank or lacks any specific reasons relating to details of the distribution	6 (22.2%)
0	Inappropriate choice on Q7ci. W(ide) = Range > 19, N(arrow) = Range < 2, H(igh) = Choices > 24, L(ow) = Choices < 26	6 (22.2%)

Table 27

The categories above are identical to those used in Q1, except that the numbers for the inappropriate subcodes have been adjusted to reflect the

different distribution of the probability situation as opposed to the sampling situation. Some sample responses include:

[L3] RL: {22, 23, 24, 26, 27, 28} While 25 flips are likely to be heads, in reality some variation is likely, so my numbers represent a range that averages 25
[L2] SX: {22, 23, 24, 25, 26, 27} They are all close to 25, ½ of 50
[L1] CS: {22, 23, 24, 25, 26, 27} It's usually not the same
[L0] SW: {18, 25, 28, 29, 34, 41 = (W)ide } Not sure

Of the subcodes, there were four (W)ide responses, one of which was also (L)ow, and one strictly (N)arrow response and one strictly (H)igh. In keeping with the potential mentioned earlier for confusion in multiple trial situations like Q7c, there were a couple of students who seemed to be thinking of sets of 100 flips, putting for example "40, 45, 50, 55, 55, 60."

Application of Q7 to Framework (Expecting): The themes relating to the dimensions of *what* and *why* were described at length in discussing Q1-Q3, and Q7 helped confirm the themes presented earlier as useful for describing responses that related to the *expecting* aspect. Some brief examples follow to show how the themes described earlier also applied to the context of probability as well as the sampling context.

Looking at *what* was expected, again responses could be seen concerning themes about the expected value, repeated values, and a range (or extreme values). Elements of the theme concerning the expected value of 25 heads include the idea that results should be or be close to the expected value, and not necessarily be the expected value each time, but results should

be on both sides of that value. In Q7ai, Table 25 shows how most students gave the expected value of 25 heads, and in Q7c the choices for six trials were at or clustered near 25. In Q7aii and Q7c, the reasons offered naturally hinged on that expected value:

(Q7aii) SX: It will be close to 25
(Q7c) BP: I chose numbers close to 25
(Q7c) MG: The average should be 25

The expected value, as MG and some other students suggest, should also serve as an average of results.

Concerning repeated values, responses in Q7b showed that most students think results should vary and not necessarily repeat. Also, in Q7c, only one student put all 25's for her choice on six trials: All the other subjects had at least some differences in values. As in Q1-Q3, responses in Q7 got me focused on how much repetition is expected or considered acceptable by subjects.

Regarding the range, again I saw some extreme values that were unlikely to occur, such as in GP's response to Q7b: "Could be 30, 25, 20, 27, if he was super super super super lucky he'd get 50." For Q7c, SP put choices on six trials that went as low as 2 heads in 50 flips of a fair coin, which is an extremely rare outcome in 6 trials.

Looking next at *why* students held their expectations, responses on Q7 reflected themes of possibilities and likelihood, variation or distribution, and proportional reasoning. For example, here are some explanations made in different parts of Q7:

(Q7aii) RL: It's [25 heads] the most likely scenario
(Q7aii) SX: The likelihood that it lands on 25 is small
(Q7c) SA: They are all possible choices

The responses above suggest a connection between the subjects' perceptions of what seems possible or likely and what they put down for their expectations. Also, some responses in Q7 specifically cited variation in their reasons, such as:

(Q7b) TO: May vary a little, but not much
(Q7c) MG: Because there should be variation around the mean

Finally, just as in Q1-Q3, proportional reasoning came through in the responses to Q7, as subjects considered the two sides of the coin to apply the 1:2 ratio to the fifty flips comprising a single trial.

Application of Q7 to Framework (Interpreting): Although Q7 offered less opportunities for responses than Q1-Q3 collectively, still all three dimensions of *causes*, *effects*, and *influencing expectations and variation* were addressed by at least some response in Q7.

For *causes*, recall how SW's response in Q7aii suggested that results may have something to do with what side of the coin was facing up to begin the flip, which implies a physical cause of variation. Two other responses that appealed to a physical cause were:

(Q7b) RL: In the absence of any change of approach, the results are likely to be the same
(Q7c) TO: It won't vary that much from [25], unless Mark alters the penny in any way

Responses such as the above point to the physics behind the coin-flipping scenario.

Regarding *effects* of variation, again the theme of reality versus probability came through:

(Q7b) DM: In all likelihood it would probably be different, but statistics say again it should be 25
(Q7c) RL: 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's and RL's have approach 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.

A second theme for *effects*, the lack of ability to make a decision, or the idea that anything can happen, showed up in responses to Q7 as subjects explained their reasoning:

(Q7b) SL: Couldn't hazard a guess, or could but it would be random
(Q7c) JM: No idea. Seemed like a nice random pattern
(Q7c) RF: This one I don't know. I have to do it physically
(Q7c) SP: Just chose randomly – Anything is possible

Again while subjects themselves may not see their thinking as an effect of variation, I view the theme of difficulty in making predictions as resulting from the variation inherent in the situation. Similarly, an effect of variation on people who don't fully appreciate what to expect is the idea that results can be anything.

Finally, the theme that arose for *influencing expectations and variation* had to do with the number of trials performed. For example, JM was cited

earlier in Q7aii for his response that appealed to the Law of Large Numbers. Some other responses talked about how they expected the average of a number of trials to be 25 heads or "about 25." The emphasis on how the probability didn't change with the number of flips or trials also came through:

(Q7aii) JL: No matter how many times he flips, the odds are the same
(Q7aii) SP: No matter how many times he flips it, the chance is still ½
(Q7c) AL: I don't see how the chances of getting heads will change if he does more sets of 50 flips

A last thought, with the opposite emphasis, came from DS in response to Q7c. She said that "the more times the coin is flipped, the more chances to deviate from 50% heads." Thus, the theme concerning the number of trials has to do with the stability of the ratio, convergence towards the expected value, and also increased possibility for variation.

Results from Q8 & Q9: These two questions were not designed explicitly to gather thoughts about variation, but as a way of seeing where subjects are in terms of probabilistic reasoning. Table 28 shows the two questions.

PreSurvey Questions #8 & 9		
Name of Task	Question Number	Description of Task
Two Spinners (50/50 & 50/50)	Q8	Angela thinks she has a 50-50 chance of winning. Do you agree? (There are checkboxes for yes/no) . Explain your answer.
Two Spinners (50/50 & 25/75)	Q9	What do you think the chances of winning this game would be? Explain your answer.

Table 28

Q8 shows two spinners, both of which are 50% black and 50% white. Q9 also

has two spinners, one of which is 50% black and 50% white, the other which is 25% black and 75% white. Both questions are about a game in which each of the two spinners is spun once, with a "win" meaning that both arrows landed on black. The categories and numbers of responses coded at each level for the two questions are given below in Table 29:

Results for Questions #8 & 9			
Question Number	Coding Level	Description of Category	No. of Students
Q8	3	[No]. Gives 1/4 or 25%, lists sample space, uses multiplication principle, or other good reasoning	15 (55.6%)
	2	[No]. Gives less than 50%, or gives 1/4 with no reason or a bad reason	2 (7.4%)
	1	[No]. Gives no reason or a poor or vague reason	1 (3.7%)
	0	[Yes]. [No]. Gives irrelevant reason or involves physical attributes of spinner	9 (33.3%)
Q9	2	Calculates 1/8 and gives reasonable explanation why	4 (14.8%)
	1	Anything less than 25%, with some reasoning	8 (29.6%)
	0	Anything greater than or equal to 25%, reasoning about the ratio of total shaded area, or uses unclear reasoning	15 (55.6%)

Table 29

As the table above shows, most of the class was able to give the correct probability for Q8, and only four subjects were able to correctly answer Q9. However, in addition to providing some baseline information about the students' ability in two-stage probability situations, some of their responses did inform the conceptual framework.

Application of Q8 & Q9 to Framework (Interpreting): The responses informing the aspect of *interpreting* were few but relevant. Although the *interpreting* aspect is about variation, and yet neither Q8 nor Q9 specifically

addressed variation, there were two themes that came through in the responses that made me think of the dimension of *causes* of variation.

One theme seemed very similar to additive reasoning, and was akin to the way that students focused on the numbers of the candies in the sampling situation as opposed to the proportion. So too in Q9, where most students did not seem to know what the correct probability was, there was some attention to the sheer area shown on the spinners:

(Q9) EM: There is more white to land on in the second spinner
(Q9) LT: There is less black space in the spinners than white
(Q9) SL: There is less black space therefore less chance of getting 2 spins in black.

The focus on the above responses is a sort of area-addition thinking, and so I wonder if in the subjects' minds there is connection – just like the numbers of candies had something to do with the chances of getting red – between the area of the white or black space and the chances of winning.

The second theme for *causes* was the physical nature of spinning. Several responses picked up on this theme, and here are a few examples:

(Q8) SW: I think it depends somewhat on where the spinner is started from and the spinner is not on the same point in both pictures.
(Q8) RF: Yes because the amount of black is the same and a lot I think depends on how you spin.
(Q9) JM: It still depends on the force, resistance, direction, and other factors.

In the future questions about variation which relied on spinner tasks, I noticed this theme again.

Now that the PreSurvey results have been summarized, I'll

share some select results from the PostSurveys. The selected questions either confirmed or added more to the framework, and for some of the parallel questions I'll be able to use the coding levels for categories of responses as a convenient way of showing some shifts in classwide results.

PostSurvey – Data & Graphs

This PostSurvey was the first one given to the class after the Math 212 class formally began the probability and statistics portion of the curriculum. Recall that the class had experience working a variety of different graph types, and the PostSurvey for Data and Graphs used several of the types seen in class. The questions for this PostSurvey were sufficiently different from those asked on the PreSurvey that instead of modifying and adapting a coding scheme to discuss results, I'll talk directly about how the responses related to the conceptual framework. There were 28 students who completed the PostSurvey (Data & Graphs).

Application of Q1ai to Framework (Interpreting): Each of the two questions on PostSurvey (Data & Graphs) had multiple parts: Question 1 focused on the differences in average monthly rainfall in Portland and Columbus over a thirty-year period (also called the Normal Rainfall). A bar chart and summary statistics were given for Q1a (see Figure 23 on the next page). Responses to Q1ai addressed the *interpreting* aspect, since the question asked students to think of causes for the different patterns of rain in the two cities.

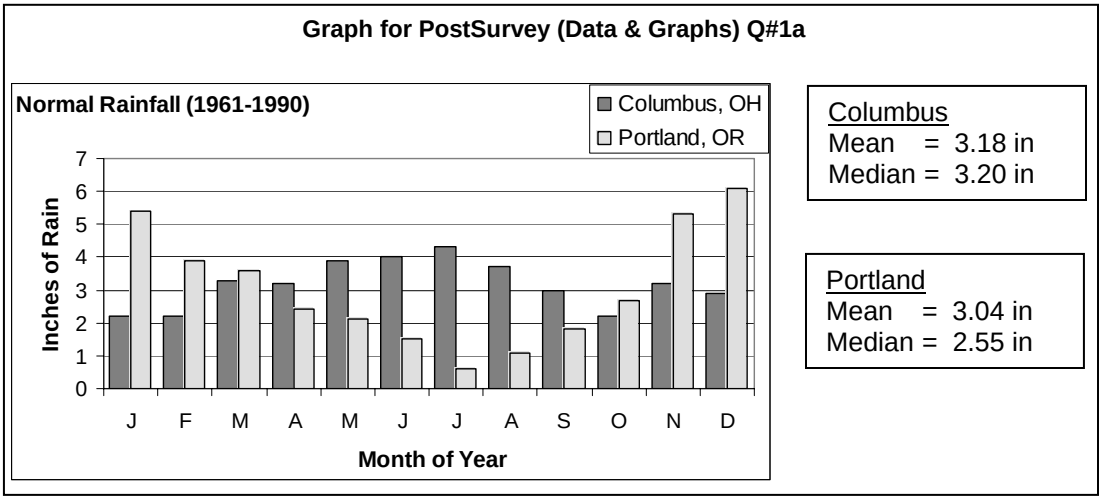


Figure 23

For the dimension of *causes of variation*, I found two themes (weather system differences and geographic location differences) which are summarized in Table 30 below.

Results for Question #1ai

Description of Theme	No. of References	No. of Students
Weather System Differences: * Snowfall * Humidity / Precipitation * Temperature * Wind Patterns * Pressure System	40	26 (92.9%)
Geographic Location Differences : [Emphasis on (P)ortland or (C)olumbus] (C) = Nearer to Great Lakes or Atlantic (P) = Near Pacific Coast (C) = Plains, Lack of Mountains (P) = Mountains or Valleys (C) = Inland, MidWest, or East (P or C) = Other features	53	23 (82.1%)

Table 30

In Table 30, the column for the number of references has to do with components of a student's response. That is, one response could contain several references. Also, a student's response could span different themes, which is why the column for the number of students exceeds the 28 total students who completed the PostSurvey (Data & Graphs). In fact, 22 students

had responses touching on both themes, and 25 students had multiple references within their responses. Examples of what two students wrote are:

> JM: Columbus gets more rainfall in the summer months. Thunder-storms 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
>
> SX: Geography of the two cities causes 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?)

Although I cannot vouch for the plausibility of all the causes that students wrote, what impressed me on this question was that every students had at least something to say about why there should be differences. No one wrote, for example, how they didn't know or how rainfall is just random and one can never tell. For whatever reasons – possibly an attenuation to weather that seems connected to the local culture of Portland – writing about *causes of variation* seemed to come easily to the class as a whole.

Application of Q1aii to Framework (Displaying): Q1aii offered a line of reasoning from a hypothetical person, Adam:

> Q1aii: Adam and Zain are discussing the data. Adam said Portland was rainier because it got the highest amount of rain a month. What do you think Adam is thinking when he says this?

The reason for the question was to get at the idea of what is meant by "rainier." Just as in class, how we had talked about "what is typical" and how the question could be approached in different ways by focusing on different

features of the graphed data, so too now I wanted subjects to reflect on the arguments that could be made for what city could be considered "rainier." However, I guided the responses by suggesting a specific line of reasoning. Recall how Figure 23 does show how Columbus has a higher mean and median for normal rainfall. However, Portland may have a reputation as a rainier city, with November and December being particularly rainy months.

Responses to this question related to *displaying* variation, and in particular to the dimension of *evaluating and comparing graphs.* I noticed the themes which are summarized in Table 31 below.

Results for Question #1aii		
Description of Theme		No. of Students
Focus on Average :	Mean or Median (or just "Average")	5 (17.9%)
Focus on Spread :	Attention to Range	2 (7.2%)
	Portland's Maximum Rainfall	6 (21.4%)
Focus on Shape or Distribution :	Compares Several Months	14 (50.0%)
	Other Distributional Reasoning	6 (21.4%)

Table 31

Again, the number of students exceeds 28 because a single response could span multiple themes, as will be the case in reporting on this PostSurvey unless otherwise indicated. The themes described in Table 31 are the same from the PreSurvey, which arose when *evaluating and comparing graphs* was discussed for PreSurvey Q5 and Q6. I split two of the themes apart to accommodate different emphases in responses which I noted. For example, with the focus on spread, some students only looked at Portland's high

month (the maximum), and a couple of students considered the entire range.

A sample response to Q1aii that included a focus on the average is:

> RL: Since Columbus has a higher average monthly rainfall, Adam is probably referring to the fact that Portland shows such extreme measurements (outliers) in several instances

RL's mention of the higher average in Columbus must involve the mean or median. The mode is not as clear from the graph, although EM spoke of Portland's December maximum of 6 inches as a mode (confounding the high bar with frequency). When RL talks about "extreme measurements" for Portland, he could be referring to both minimum and maximum, but a different student wrote more directly about the range:

> SX: Portland's highest rain months have much more rain than Columbus' most rainy months. Adam and Zain must be looking at the ranges and the range for Portland is much higher than the range for Columbus.

Whereas SX looked at the entire range, another student, SW, just looked at the maximal value for Portland and nothing else in making her response (as did some other students):

> SW: I think Adam is looking at the graph and he saw that in the month of December, Portland got 6 inches of rain which is the most inches of any month.

As Table 24 shows, half of the students used several months in their comparison, which seemed to me to reflect focus on distribution: That is, they did not rely on just one attribute or characteristic of the graph, but instead they consider a cluster of data, typically the highest three months for Portland:

CS: He sees that in January, November, and December has the most rain for Portland and that Ohio never reaches any of them.

JL: He means Portland had the significantly higher amount of rain in November, December, and January. Those three months fill the outer range on annual rain.

Some students had responses that included a focus on distributional reasoning in addition to other themes, such as SZ:

SZ: Adam probably just took a quick look at the data and focused on five bars representing Portland's rain amount. Those bars are on the outside edges of each side, spread out from each other, but clustered together: Columbus' is spread out throughout the middle.

Notice how SZ's language invokes the distribution, with phrases like "clustered together" and "spread out through the middle." I was impressed at some of the synthesis of ideas that a few students used in their responses, bringing together ideas of average, spread, or distribution.

Application of Q1bi to Framework (Displaying): Q1b used the same data as Q1a, except that I included two boxplots showing the data. Students could, of course, refer back to the earlier page showing the bar charts and summary statistics. The question for Q1bi offered a different line of reasoning from Q1aii:

Q1bi: Zain says Columbus was rainier because the average monthly rainfall was higher than Portland. What do you think Zain was thinking when he said this?

Like Q1aii, the responses to Q1bi also got at the *evaluating and comparing graphs* dimension of the *displaying* aspect. Since the question naturally leads to a focus on average, it is no surprise that most of the responses did

exactly that. However, there were a few responses that touched on the other two themes, and Table 32 summarizes the responses:

Results for Question #1bi	
Description of Theme	No. of Students
Focus on Average : Mean or Median (or just "Average")	21 (75.0%)
Focus on Spread : Attention to Range or IQR	5 (17.9%)
Focus on Shape or Distribution :	5 (17.9%)

Table 32

Because boxplots were available, many of the responses focusing on the average involved the median, as EM wrote: "Zain was thinking about the median and noticing that it is much higher than Portland's mean and median." Notice how EM compares two types of average in her response. Also, in focusing on spread, some responses referred to the interquartile range (IQR), like JL's response: "He means the Interquartile range for monthly rainfall is greater in Portland than Columbus." Finally, some responses included language that reflected distributional thinking, using phrases such as "a constant concentration of rain," or how the "rainfall was condensed during the months." Two sample responses in their entirety are:

RF: 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 it looks more compact, pretty much it is almost the same rain all year.

RL: Zain is ignoring the range of each city's rainfall measurements, as well as the extent to which Portland's rainy months are rainy. He avoids the issue of distribution.

What I particularly noticed in Q1bi was not only the application of the

framework to the responses, but how the boxplots seemed to give students another avenue for talking about spread and distribution. In class, there had been ideas discussed about the IQR and its relationship to how data clustered together, and those ideas seem to come through in some of the responses.

Application of Q1bii to Framework (Interpreting and Displaying): Having led the students in their responses by providing lines of reasoning from the hypothetical Adam and Zain for them to react to, Q1bii just asked: "What city do *you* think is rainier, and why?" It is true that the question allows for subjective reasons - and 9 students included their opinions - but responses also included other themes which I've identified in Table 33.

Results for Question #1bii		
Description of Theme	Aspect / Dimension	No. of Students
Causes of Variation: Reasons for the differences	Interpreting / Causes	7 (25.0%)
Subjective Ideas: Beliefs or feelings or attitudes	Displaying / Making Conclusions	9 (31.1%)
Focus on Average: Mean or Median (or just "Average")	Displaying / Evaluating & Comparing	11 (75.0%)
Focus on Spread: Attention to Range or IQR	Displaying / Evaluating & Comparing	12 (17.9%)
Focus on Shape or Distribution :	Displaying / Evaluating & Comparing	6 (17.9%)

Table 33

As Table 33 shows, some students did bring reasons for the differences into their responses, and there were also students who blended other elements (such as subjective ideas, or attending to average, spread, or distribution) together in making their final conclusion. The table does not show the final conclusions of the 28 students: 11 for Portland as the rainier city, and 17

for Columbus. A sample of some of the responses will show the breadth of reasoning:

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

EM: Columbus is more consistently rainy by looking at the boxplot. It's interquartile is smaller and reflects less change.

SA: Portland, because I live here and it rains all of the time. Really I think Columbus has more rain because the amount of rain they get every month is consistent. In Portland we have really low & really high months.

LT: Who gets more rain on average is Ohio. The mean and median are both higher than Oregon.

In the responses above, all the themes other than causes (which were discussed previously) can be seen: There are some subjective ideas, from the tacit notion of DM's "massive rains" to the blatant comment of SA. There is also a focus on average from LT, and a mention of the IQR by DM and EM. Finally, elements of distributional thinking show in the responses of DM ("more steady") and EM ("more consistently rainy") and SA ("rain…is consistent"). Although the term "consistent" in this task appeals to the way the graph for Columbus looks, and does reflect distributional thinking, I came to wonder if responses that emphasized consistency also might have something to do with the dimension of making conclusions.

Q1a and Q1b offered many opportunities to evaluate and compare graphs, and Q1bii in particular brought out some multi-faceted responses. Collectively, the whole class reasoned on Q1bii using many characteristics of the distribution, and I could see from what they wrote that the language of the boxplots was useful. I was surprised at how many (11 students) people

wrote that Portland was rainier, because most of those 11 students did not argue on the basis of subjective reasoning but instead offered a statistical justification.

Application of Q1c to Framework (Displaying): This was the second of three opportunities in the survey instruments that students had to *produce* a graph, and the questions read as follows:

> Q1c: In Columbus, the normal monthly rainfall for the month of June is reported as 4 inches. Draw a graph below which shows how many inches of rain Columbus might get for each day in June.

Axes were provided, with the days of June on the horizontal and inches of rain on the vertical axis.

Within the dimension of *producing graphs*, I regarded the graphs along the same themes that arose in the PreSurvey, and I summarized the themes as shown in Table 34 on the next page. Two of the 28 students turning in the PostSurvey (Data & Graphs) left Q1c blank, so the percentages shown are out of the total 26 respondents. As in the PreSurvey, the parts of the theme concerning technical details of graph-making included the type of graph used and the configuring of the scales along the axis, and the parts of the theme concerning the characteristics of the distribution included the reasonableness of the average, spread, and shape. Some of what constituted a reasonable shape involved a judgment call: For example, some graph had many days with no rain, which I considered reasonable in considering what might happen during a month.

Results for Question #1c			
Description of Theme			No. of Students
Technical Details	Type of Graph	Bar Chart (or Similar)	11 (42.3%)
		Line Plot	7 (26.9%)
		Straight Line	4 (15.4%)
		Other (ie, Unconnected Dots)	4 (15.4%)
	Scale on Axis	Adequate	14 (53.8%)
		Inadequate	12 (46.2%)
Characteristics of Distribution	Average	Correct	12 (46.2%)
		Incorrect	14 (53.8%)
	Spread	Appropriate	21 (80.8%)
		Inappropriate	5 (19.2%)
	Shape	Reasonable	15 (57.7%)
		Unreasonable	11 (42.3%)

Table 34

Several things stand out in the Table: The most common graph type was a bar chart, and most people had an adequate scale along the vertical axis. For Q1c, 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. Since the vertical axis provided was marked “Inches Per Day”, it would be inappropriate to put, for instance, a horizontal line at 4 inches (as some students did). In other words, even if it did rain 4 inches every day in June, that does not translate to a 4 inch average monthly rainfall. Thus, where I have put “Average” and “Correct / Incorrect” in the table above, I am referring to students who calculated the daily average of 0.13 inches. The fact that most students had an incorrect average (typically a daily average of 4 inches) is what threw many graphs off. However, I tried to consider whatever daily average they had as valid in looking at their graphs for spread and shape.

Some examples of their graphs would be useful to see at this point, and I'll give two each for the following types: Bar chart, line plot, and straight line. Here are DP and BP's bar charts:

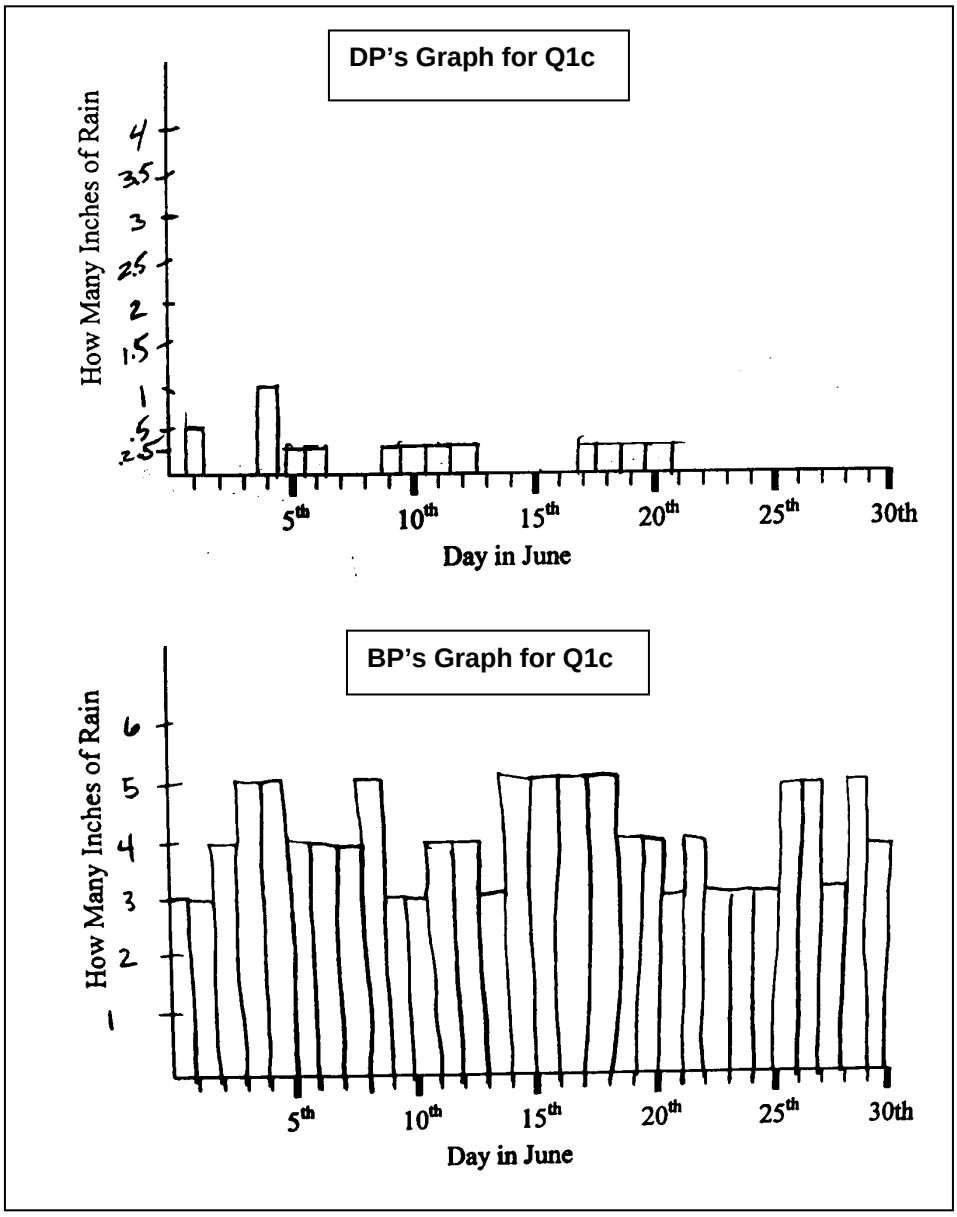


Figure 24

Notice how DP's bars are chosen so that they easily add up to 4 inches in one month. BP's graph is an example which I thought had reasonable

variation (shape and spread), assuming the daily average was actually 4 inches. In other words, the day-to-day fluctuations for rainfall that BP was trying to show in her graph was plausible. Here are MA and DS's line plots:

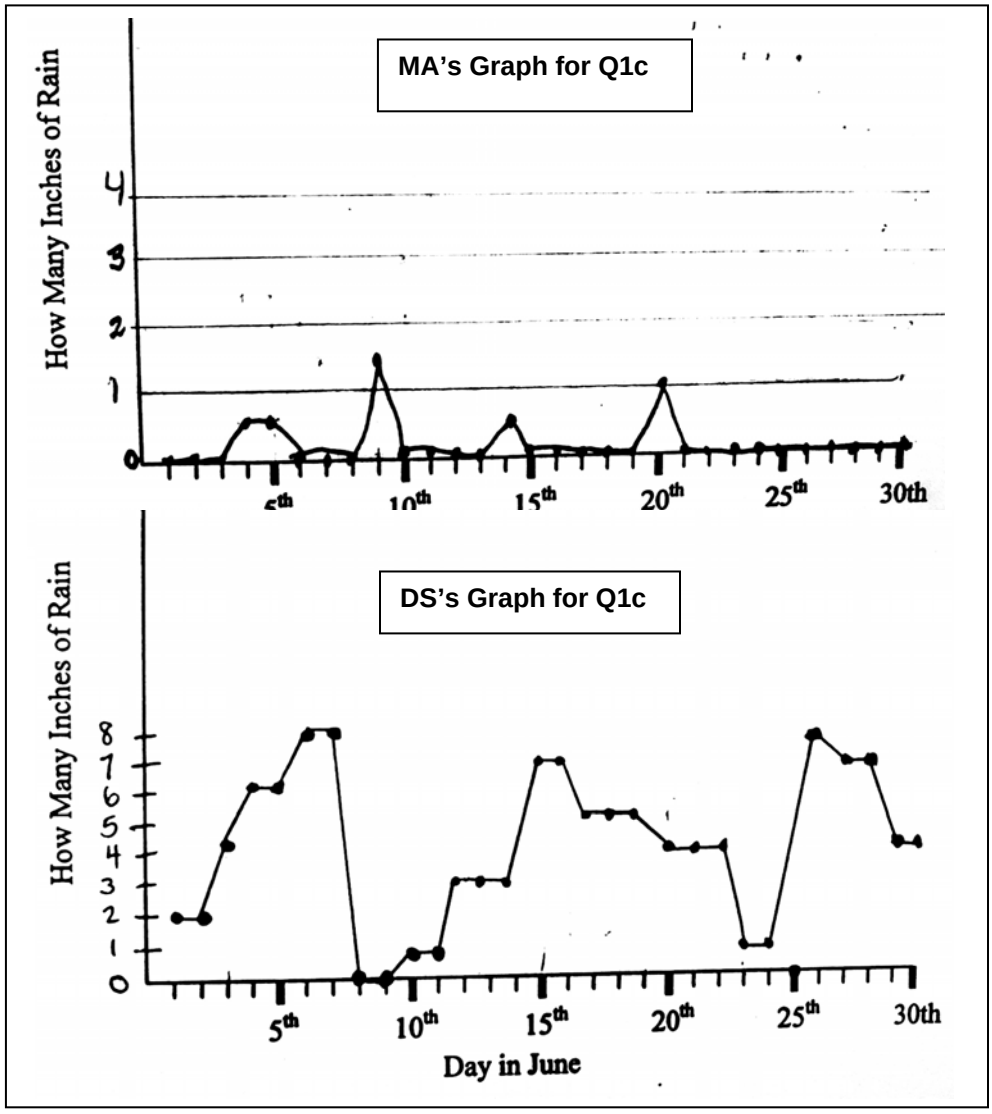


Figure 25

Again, in MA's graph the days with rain are conveniently chosen: From left to right, she has days with 0.5", 0.5", 1.5", 0.5", and 1.0" of rain for a monthly total of 4". With DS's graph, again I thought she gave a good

example of day-to-day variation. Even though her daily average was not correctly calculated, her graph would be reasonable under the assumption that a daily average of 4" was called for. Finally here are two graphs with inappropriate variation shown – the straight lines of RB and RL:

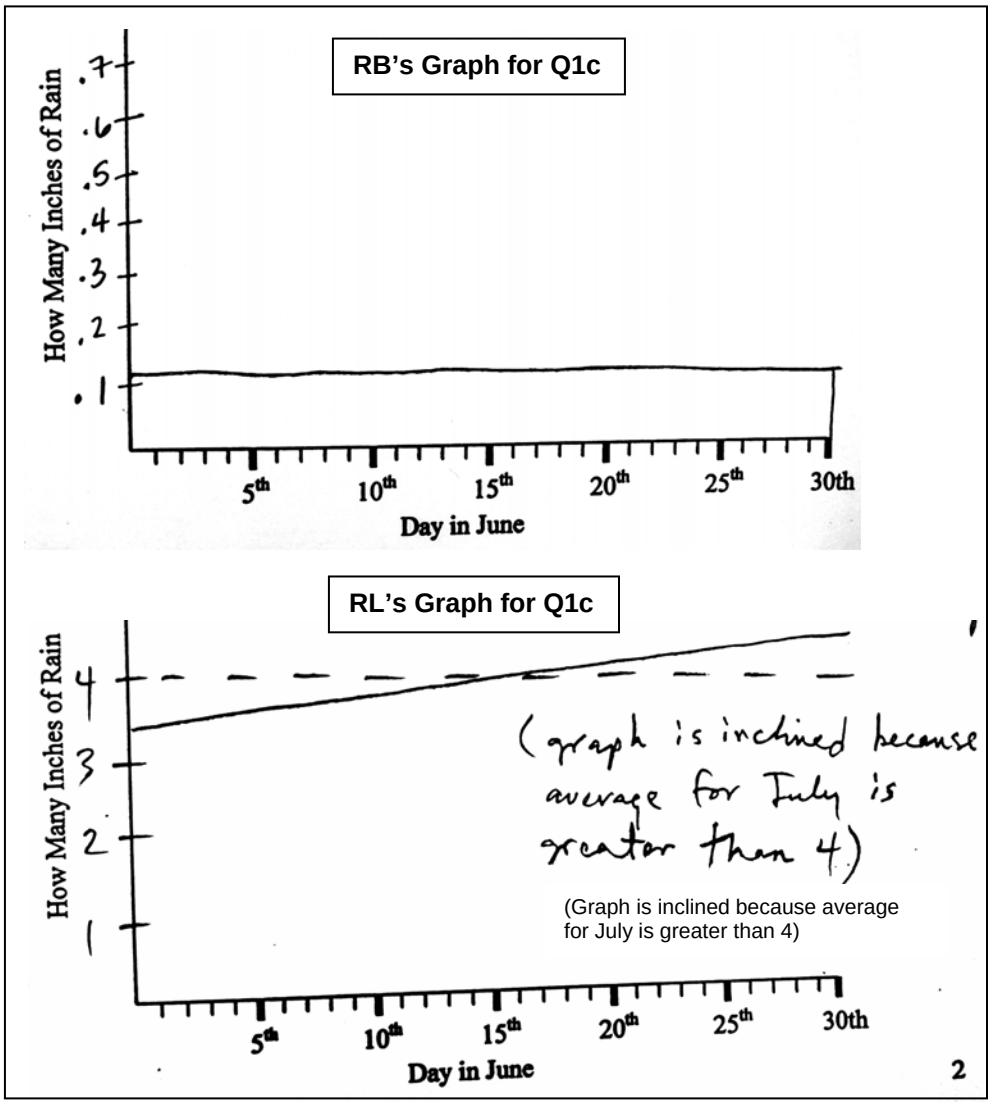


Figure 26

Even though finding the correct daily average is one approach to graphing the situation called for in Q1c, RB above did have notation showing 0.13" average daily rain, but she has implied with her graph that every day in

June has exactly the same rainfall, with no variation. RL shows his dashed line where the supposed average daily rainfall of 4" would be, and shows (instead of unpredictable rainfall) a steady increase in daily rain.

What I learned from the graphs of Q1c is that the themes for the dimension of *producing graphs*, along with the elements of the themes shown in Table 27, are useful for looking at the variety of ways that the students had of trying to graph the situation for average rainfall. It is true that some confounding issues were making sense of the monthly average versus the daily rainfall, but I could see that most graphs tried to convey a sense of day-to-day fluctuations, or even stretches of no rain followed by appropriate amounts of occasional rain. It clearly is not a trivial task coming up with graphs for many students, let alone graphs that attend to technical details as well as reasonable characteristics of the distribution.

On the PostSurvey (Data & Graphs), there was also a Q2a and Q2b which were similar to Q1a and Q1b in the sense that students were asked to compare graphs. For Q2 the graphs were two dotplots and two boxplots, and the context was about different traffic death rates in different parts of the country. I'm not reporting on the results for this question because although they were interesting, the themes were very similar to those already illustrated (and which will be illustrated more in the PreInterview data and PostInterview data). Instead, I'll move on to showing results from the second PostSurvey, which was in the context of sampling.

PostSurvey - Sampling

The PostSurvey for the Sampling context was handed out after we had done the Known and Unknown Mixture activities in class. There were four questions on the PostSurvey (Sampling), with the first question (Q1) using the same sampling scenario as was used on the PreSurvey: Handfuls of 10 candies drawn from a jar containing 60 red and 40 yellow candies. I later called the jar in Q1 the Small Jar to distinguish it from the Large Jar used in Q2 – Q4. For sampling from the Large Jar, handfuls of 100 were pulled from a mixture of 600 red and 400 yellow candies. Because Q2 – Q4 on the PostSurvey (Sampling) were isomorphic to Q1 – Q4 on the PreSurvey, I'll discuss them separately from Q1 later in this section. All thirty students in Steve's section completed the PostSurvey (Sampling).

Application of Q1 to Framework (Expecting and Interpreting): There were two parts of Q1, and I will report on the second part (Q1b) because it added a new theme to the dimension of *why* students held their expectations. Responses to Q1b also had some different emphases on previous themes, I suspect because the nature of the expectation for Q1b was a sort of reversal from the kinds of questions asked earlier. Q1a and Q1b both related to a boxplot that showed twenty trials at the Small Trials which had really happened (that is, actual data was used from a computer simulation). The twenty trials (which corresponded to twenty people doing one trial each) attained a minimum of 3 reds in a handful of 10. Thus, Q1b read:

Mike was surprised that nobody got 0 or 1 Red candy in their handful. He decides that he's going to try it with more than 20 people.
(Q1bi) How many people do you think Mike should have do this so that at least one person gets 0 or 1 Red candy in their handful ?
(Q1bii) How did you decide on that answer?

Logically, Q1bi asks for an expectation of how many trials should be run until results hit less than or equal to 1 red candy. I worded the question in terms of getting 0 or 1 red candy because this was a take-home item and I wanted the question to be as clear as possible: A "success" for Mike will happen if he sees either a 0 result or a 1 result, and the wording of the question was an attempt to take into consideration that a 0 red could happen before a 1 red.

I grouped responses to Q1bi in the categories shown below in Table 35:

Results for Questions #1bi	
Description of Category	No. of Students
Suggests over 10000	1 (3.3%)
Suggests 1000 to 9999 people	5 (16.7%)
Suggests 100 to 999 people	16 (53.3%)
Suggests 0 to 99 people	4 (13.3%)
Irrelevant answer or No Answer	4 (13.3%)

Table 35

While it is possible to get a result of 0 or 1 red in only few trials, what I considered reasonable for students to write as an estimate for the number of people (trials) in Q1bi would be at least a few hundred. In the category "100 to 999" above, 7 of the 16 students either put 100 trials or suggested that Mike start at 100 and go up from there. Instead of raw calculations, which weren't necessary for this question, my judgments stemmed from the trials we had

done in class and also the time we spent looking at the ProbSim software doing hundreds and thousands of trials.

The classwide experience was exactly the new theme that emerged for the *why* dimension of the *expecting* aspect. The responses for Q1bii (and a few responses that came from Q1bi where students added commentary to their expectations) are summarized in Table 36 below, along with the aspect and dimension to which the theme was attached. Some samples responses for each theme follow.

Results for Question #1bii		
Description of Theme	Aspect / Dimension	No. of Students
Involves Proportional Reasoning	Expecting / Why	9 (30.0%)
Involves Possibilities and Likelihoods	Expecting / Why	11 (36.7%)
Involves Experiential Reasoning	Expecting / Why	13 (43.3%)
Number of Trials	Interpreting / Influencing Expectation & Variation	9 (30.0%)
Number of Candies	Interpreting / Influencing Expectation & Variation	5 (16.7%)

Table 36

For proportional reasoning, because the nature of the expectation is not as straightforward as just asking what would likely be in a handful of 10, the responses that included a specific mention of some kind of proportional thinking were more complex:

CS: Well, I looked at 60 Red and 40 Yellow and thought of the odds as 6 – 4, and the median is 6. So then I multiplied 60 x 6 = 360.

RL: I multiplied the chance of getting one red (6/10) times ITSELF once for each red the question asked for. {He has written in the margin: “1 Red = (6/10)*(4/10)^9=(1572864/1000000000) ~ 16/10000”}

It's not quite clear to me why CS performed the calculation she did, but I can see RL's line of reasoning about multiplying individual probabilities. Some other students reasoned about *why* not by using proportions, but by using the median as a measure of center, and I began to think that the theme of proportional reasoning might be connected to reasoning using centers in general.

The theme involving possibilities and likelihoods again emerged in the responses, including language of probability, chance, and what could or could not happen:

LW: The odds are very unlikely that someone will pull 0 or 1 red candy.
SX: Because the likelihood is so small that only 0 or 1 red candy would be pulled
SZ There are 60% red candies compared to the 40% yellow so…the possibility is greater that more red will be pulled out
SA: It's impossible to get zero red.

Except for SA, the language in the theme of possibility and likelihoods offered by the students tends to be somewhat subjective, with phrases like "very likely" leaving open the question of how likely that would be exactly. I wondered if SA's strong statement of the impossibility of getting a 0 result is a reaction to the time we spent looking at multiple trials on the computer without obtaining that result. Also, SA might be writing "impossible" yet really meaning "almost impossible," as EM wrote: "After seeing the simulations in class on the computer, it seemed almost impossible to get a zero."

EM mentions the in-class experience, and the theme involving

experiential reasoning (in this case having to do with the class experiences) came through in 13 student responses (the highest number for any given theme in Table 36). I'm giving more than a few examples to really offer a good sample of what students said for this new theme:

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
GP: I just know from doing classroom work that it would have to be pretty high
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.

When SA mention how "we saw it…in class", she does not mean that we used the same mixture, because on the PostSurvey (Sampling) the mixture was 60 Red to 40 Yellow, and in class the mixture was isomorphic to 70 Red to 30 Yellow. One student picked up on the different mixtures:

DS: Odds are greatly against getting only 0 or 1 reds, and when we did the test on the computer it took 5000 to get 0 or 1 when there were 70% reds. I lowered the # to 1000 because there are 60% reds in this scenario.

It was clear to me that the class experience had made an impression on student thinking, and I made this theme a part of the *why* dimension of the *expecting* aspect because it seemed that students were offering what they had seen (experienced) in class as a justification for why they had put the expectations they did in Q1bi.

The last two themes in Table 36 had to do with the *interpreting* aspect, and both themes addressed the dimension of *influencing expectation and*

variation. The first theme, involving the number of trials, came through as students reasoned explicitly about the premise of this question, which was that more trials would eventually produce the unlikely results:

AL: You need more tries for the spread to be greater
CM: The more people he has pull a sample of 10, the more likely it is that someone will pull 0 or 1 red
DP: The more tests you do, the greater chance of getting an outlier of 0 or 1

Whereas DP phrases her response in terms of a "greater chance" resulting from more tests (trials), other reiterated the stability of the underlying proportion, communicating how the ratio doesn't change no matter how many trials are performed. RB expresses 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." Thus, the theme involving the number of trials spans a variety of idea: More trials increases the spread, more trials gives more chances to obtain outliers, and more trials has no effect on the underlying ratio.

The second theme for the *interpreting* aspect involved the number of candies in the jar, which we had seen in some responses on the PreSurvey as a kind of additive reasoning. For those students who seem to focus on the "more reds" kind of thinking, it does not seem that the proportion makes as much difference to their perception of the situation as does the sheer number of candies.

SA: I'm not sure why it's pretty much impossible to get just 1 red or zero reds, but 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.

Whereas Q1b has been described in terms of themes of the framework, Q2 – Q4 will be portrayed first in terms of the same kinds of categories and coding scheme as was used in the PreSurvey, because the questions are completely parallel. After describing the categories and coding levels, again the results will be briefly related to the framework.

Results for Q2-Q4: A comparison of Table 37 on the next page with Table 11 earlier in this chapter will show that PostSurvey (Sampling) Q2 – Q4 are isomorphic to PreSurvey Q1 – Q3, and the tasks even have the same or similar names. In the PreSurvey, the tasks were based on samples of 10 from the Small Jar; with the PostSurvey (Sampling), the tasks were based on samples of 100 from the Large Jar. One difference in the questions is that on the PostSurvey (Sampling), after having the students make a guess for the results of one trial (Q2a), I did not have them explain their reasoning as I had done in the PreSurvey. Another difference in questions is that on the PreSurvey, the expectation on ranges started with 6 trials (PreSurvey Q2a) and then went up to 30 trials (PreSurvey Q2b). For the PostSurvey (Sampling), the similar question started with a range for 30 trials (Q3a) and then went up to 300 trials (Q3b). However, the main point of both questions was the same, which was to get at the idea of increasing numbers of trials leading to an expansion of the range.

PostSurvey (Sampling) Questions #2-4		
Name of Task	Question Number	Description of Task
One Trial	Q2a	You do one trial. How many reds do you think you might get?
Several Trials	Q2b	You do several trials. (Q2bi) Would this many reds come out every time? (Q2bii) Why?
Six Trials	Q2c	Six people do trials.(Q2ci) Write down how many reds they might get. (Q2cii) Why did you choose those numbers?
Range 30	Q3a	In 30 trials, what might the numbers of reds go from (low to high)?
Range 300	Q3b	In 300 trials, what might the numbers of reds go from (low to high)?
Why on Ranges	Q3c	Why do you think this?
Fifty Trials	Q4a	Out of fifty trials, how many would have 0-10 Red, 11-20 Red …91-100 Red?
	Q4b	Why do you think the numbers you wrote are reasonable?

Table 37

The categories used were the same as those used in the PreSurvey, with appropriate number adjustments made to account for the larger size of jar and the new expected value of 60 reds in a sample of size 100. Again some judgment calls were made in determining the parameters on some of the categories.

Table 38 on the next page begins the summary of the categories of responses as well as the numbers of students who responses were coded at each of the levels. Because the categories are isomorphic to those in the PreSurvey, and examples have already been given to illustrate the meaning underlying the categories, I have streamlined the presentation of these PostSurvey (Sampling) results to include the all the sub-parts of a question in a single table. That is, Table 38 gives all the parts of Q2, and Table 39 gives all the parts of Q3, etcetera. A similar treatment will follow for the PostSurvey

results in the context of probability, after which some classwide comparisons will be made on the basis of the coding levels used with the survey data.

Results for Question #2			
Question Number	Coding Level	Description of Category	No. of Students
Q2a	2	Either gives a range around 60, such as 50-70, or else writes (for example) "Around 60"	7 (23.3%)
	1	Gives 60 Reds as answer	13 (43.3%)
	0	Gives one number other than 60 Reds, such as 40 Reds	10 (33.3%)
Q2b	3	Q2bi = No & Q2bii = Explicit reasoning, using both proportional thinking and variation	12 (40.0%)
	2	Q2bi = No & Q2bii = Some indication of variation, supported by vague reasoning	10 (33.3%)
	1	Q2bi = No & Q2bii = No reason, or vague or irrelevant reason. Q2bi = Yes & Q2bii = Reasoning explicitly acknowledges variation	7 (23.3%)
	0	Q2bi = Yes & Q2bii = No acknowledgement of variation	1 (3.3%)
Q2c	3	Appropriate choice on Q2ci & Explanation explicitly involves proportional reasoning as well as variation	8 (26.7%)
	2	Appropriate choice on Q2ci & Explanation reflects proportional reasoning or notions of spread	5 (16.7%)
	1	Appropriate choice on Q2ci & Explanation left blank or lacks any specific reasons relating to details of the distribution	5 (16.7%)
	0	Inappropriate choice on Q2ci. W(ide) = Range > 20, N(arrow) =Range < 5, H(igh) = If (Min>60) or (Max>75), L(ow) = If (Min<45) or (Max<60)	12 (40.0%)

Table 38

On Q2a (One Trial), more students gave the expected value (60 reds) as opposed to any other kind of answer. I wondered if some of the responses coded at Level 0 were actually well aware of the expected value but deliberately wanted to suggest a different value because they suspected results might vary. Since I did not have the students put a reason for their expectation, that is a missed opportunity for gathering information on *why*; however, I was able to ask *why* on the exact same question in the PostInterviews. For Q2b, most of the students had responses coded at the

top two levels, showing at least some indication of variation or offering explicit reasoning involving both proportional reasoning and variation. With Q2c, the table does not show how the inappropriate choices broke down by subcode. There were 4 responses strictly (W)ide, 3 strictly (H)igh, and 1 strictly (N)arrow, with the narrow one being all 60s. Furthermore, 3 responses were both (W)ide and (L)ow, and 1 was both (W)ide and (H)igh. Since most of the inappropriate choices had the characteristics of being (W)ide, here is an example of a (W)ide choice and reason:

> [L0] SL: {30, 60, 60, 65, 70, 80} I chose random numbers. But I think it would tend to be higher numbers, 50 or above, since there are more reds than yellow.

Many responses on the range tasks of Q3 also were inappropriate, as Table 39 shows:

Results for Question #3			
Question Number	Coding Level	Description of Category	No. of Students
Q3a	1	Appropriate Choice: Min (45 to 55) – Max (65 to 75)	6 (20.0%)
	0	Inappropriate Choice: W(ide) = Range > 30, N(arrow) = Range <20, H(igh) = Min > 55 or Max > 75, L(ow) = Min < 45 or Max < 65	24 (80.0%)
Q3b	1	Appropriate Choice: Min (40 to 50) – Max (70 to 80)	6 (20.0%)
	0	Inappropriate Choice: W(ide) = Range > 40, N(arrow) = Range <20, H(igh) = Min > 50 or Max > 80, L(ow) = Min < 40 or Max < 70	24 (80.0%)
Q3c	3	Explicit mention of increased # of trials leading to increased range, with additional details describing the distribution	6 (20.0%)
	2	Explicit mention of increased # of trials leading to increased range (Note: Could be in the form of mentioning class experience)	15 (50.0%)
	1	Reasoning about one or both ends of the range without explicitly tying the results to the increasing number of trials	8 (26.7%)
	0	No answer or irrelevant answer	1 (3.3%)

Table 39

Again, although Table 39 shows that most of the class put inappropriate

choices on both range tasks, the table does not show that 17 responses were (W)ide in Q3a and 19 responses were (W)ide in Q3b. What seems to have happened is that choices for low to high looked very much like they did on the PreSurvey, only multiplied by a factor of 10. So, for instance, if a range for 30 trials at the Small Jar (as in the PreSurvey) went from 3 to 9, then a range for 30 at the Large Jar (as in the PostSurvey for Sampling) went from 30 to 90. Then, if a student was operating under the idea that more trials expanded the range, the range for 300 trials would be even wider than for 30 trials. The net effect tended to be that both ranges were wide, and I suspect that the relationship of population and sample size to the expected variation was not clear to most students. However, as Q3c shows, most of the students had responses at the top two coding levels for reasoning, meaning that they knew about the relationship between number of trials and the expected range. GP, whose choices on Q3a (40 to 80) and Q3b (38 to 95) were both wide, said in his Level 2 response: “The range should increase the more you do it. More chances to hit the higher and lower numbers.” JM also had wide choices (38 to 87 on Q3a, and 28 to 92 on Q3b), and his Level 3 response was:

> JM: We know that 0 reds and 100 reds are possible, but highly unlikely. The mix is 60% reds, so with 300 trials we should get a good range of outcomes. 92 reds may be high, but in 300 trials perhaps not too unlikely.

GP and JM shows plausible reasoning, despite having made wide choices.

Finally, for Q4, in picking frequencies for 50 trials at the Large Jar I had already predetermined the bins where frequencies would be gathered:

0 – 10 Reds, 11-20 Reds, etcetera. The modal bins would be expected at 51-60 Reds and 61-70 Reds. As in the PreSurvey, I analyzed their frequency charts according to the reasonableness of three characteristics:

(M)ode: Should be at Bin 51-60 or Bin 61-70
(E)xtremes: Should not be more than 5 trials at Bin 41-50 or Bin 71-80
(D)istribution: Should not be too uniform, or too skewed, or too clustered (i.e. Have a modal frequency higher than 35)

Again, there is subjectivity involved in deciding what characteristics are reasonable, and I made my decisions after considering statistical predictions for 50 trials and also after examining multiple simulations of 50 trials using the Fathom software. Table 40 gives the categories and results for Q3a:

Results for Question #4a

Coding Level	Description of Category	Number of Students
3	All three characteristics are reasonable	1 (3.3%)
2	Exactly two of the three characteristics are reasonable	22 (73.3%)
1	Only one of the three characteristics is reasonable	7 (23.3%)
0	No answer, or none of the three characteristics is reasonable	0 (0.0%)

Table 40

As Table 40 shows, most of the class had at least one unreasonable characteristic. The characteristic that was most often unreasonable was the (E)xtremes, with 27 of the choices for “Fifty Trials” being wide. That is, the choices included too many low values, and too many high values. Interestingly, not one student left this blank or put completely unreasonable guesses for the fifty trials.

In explaining reasons for their choices (Q4b), the categories are

similar to those used in earlier questions, and are presented along with results in Table 41:

Results for Question #4b		
Coding Level	Description of Category	No. of Students
3	Mentions the shape of the distribution, or uses proportional reasoning with some explicit statement about spread (and possibly refers to class experience)	9 (30.0%)
2	Uses proportional reasoning (for example: ratio, average, or percent) or vague reference to spread, or refers to class experience	12 (40.0%)
1	Uses additive reasoning (that is, the number of candies in the jar)	7 (23.3%)
0	No reason, a vague reason, or an irrelevant reason	2 (6.7%)

Table 41

Although comparisons from the PreSurvey to the PostSurveys will be made shortly, there are two differences between the results shown in Table 41 (PostSurvey (Sampling) Q4b) and the parallel results back from Table 18 (PreSurvey Q3b) that are important to bring up now. One difference is that I had included the "additive reasoning" category (Level 1) in the earlier Table 18 even though no responses were coded at that level, and here in Table 41 it can be seen that some students actually did have responses reflective of additive reasoning. The second difference is that I added the theme of class experience to the categories at the top two levels, because as Table 41 involved a PostSurvey, some responses mentioned the class experience in their reasoning.

Application of Q2-Q4 to Framework (Expecting and Interpreting): Some of the same dimensions and themes that came from responses to questions in the PreSurvey also showed in the responses to the parallel questions on the

PostSurvey for sampling. Fewer examples are provided below, but enough to briefly illustrate some of the themes that arose, themes which have already been discussed.

For instance, regarding *what* was expected, again responses had themes concerning the expected value (see SA's response below), whether or not values might repeat (see SX) , and what kind of range or extreme results might occur (see JL and BP):

(Q2bii) SA: I think the mean would be around 60 (for reds), but there would also be other numbers higher and lower than 60
(Q2bii) SX: Because there is variation – Few handfuls would be exactly the same
(Q2bii) JL: You will get 58-64 reds every time, give or take 60% reds
(Q2bii) BP: There is a CHANCE to pull out anywhere from 0 to 100 reds

Concerning the *why* behind their explanations, some responses had a theme involving the language of possibilities and likelihoods (see AL below) , or involving proportional reasoning (see DP), or involving variation or the distribution (see SX and MA).

(Q2bii) AL: This [60 Reds] is the most likely probability, but it's not guaranteed.
(Q2bii) DP: Since the ratio is similar 60/40 : 600/400
(Q2cii) SX: They are all numbers close to 60 but all different to account for variation.
(Q2cii) MA: They are clustered within what would be the interquartile range, had the process been graphed. The greatest concentration would be right around and including 60

The new theme for *why* which came out of this PostSurvey had to do with experiential reasoning, as discussed in PostSurvey (Sampling) Q1, and here too in Q2-Q4 some additional responses related to the theme of class

experience:

(Q4b) EM: From a similar exercise in class, we found that most colors were pulled around the median number with just a few outliers

(Q4b) SR: Recalling a similar activity in class, I thought the majority would get from 31-60, but some would get as low as 11-20 and as high as 81-90.

In the *interpreting* aspect, there were again responses for the theme connecting the physical environment of the sampling situation with a *cause* of variation. Also, for an *effect* of variation, there was the theme of how what probability suggests does not always match with what really happens. Instead of having students comment on the difficulty of making a decision, students readily made choices but sometimes had responses noting how anything was possible – A theme I had included under the *effects* of variation. Often the choices and reasons were quite good, but it seemed as if the student had a need to state for the record that anything could happen. For *influencing expectation and variation*, there were many responses (as Table 39 shows) supporting the theme connecting the number of trials to an expanding range. The theme of the number of candies also shows up in the category of additive reasoning in Table 41. I'll share some sample responses from the PostSurvey (Sampling) since the related Table 18 for the Small Jar did not have any responses coded at that same level:

(Q4b) LW: Since there are more red than yellow I believe it more likely for the trend to push higher rather than lower.

(Q4b) MG: Because there are so many more red than yellow, they will be more likely to pull more than 60 rather than less

(Q4b) GP: Since there are more red choices, it should skew toward the lower numbers.

The aspects, dimensions, and themes of the framework were reinforced by responses in the PostSurvey for the context of sampling, and what I noticed was that the overall depth of the responses had increased from the PreSurvey. The students seemed to include more language that appealed to the distribution, talking about how results “clustered” or were “spread”, and I suspect that having more vocabulary to use played a part. For example, by the time of the PostSurvey (Sampling), we had worked with a variety of graph types in class, noting how different graphs portrayed data, and we had also explicitly discussed variation in the sampling context (using many different descriptive phrases). The trend toward fuller explanations that explicitly related to distributional reasoning continued into the PostSurvey results for the context of probability.

PostSurvey - Probability

The PostSurvey for the context of Probability was given as Steve’s class was finishing with the stochastics part of the course and making a transition back to geometry for the last two weeks of the quarter. The PostSurvey (Probability) was completed by 29 students, after all classroom activities and virtually all discussion having to do with probability and statistics had already taken place.

There were three questions on PostSurvey (Probability), and all three were similar to earlier questions asked on the PreSurvey. Because of the

similarity in questions the PostSurvey (Probability) questions will be portrayed first in terms of the same kinds of categories and coding scheme as was used in the PreSurvey, After describing the categories and coding levels, again the results will be briefly related to the framework.

Throughout the PostSurvey (Probability), one trial consisted of spinning a half-white and half-black spinner fifty times, and noting how many of the fifty spins landed on black. The questions themselves are listed below in Table 42:

PostSurvey-Probability Questions #1-3		
Name of Task	Question Number	Description of Task
One Trial	Q1ai	Matt does one trial. How many blacks do you think he might get?
	Q1aii	Why do you think this?
Compare Trials	Q1b	Matt does a second trial. (Q1bi) How do his results on the second trial compare with the results on the first trial? (Q1bii) Why?
Six Trials	Q1c	Matt then does six trials. (Q1ci) Write down how many blacks he might get on each trial. (Q1cii) Why did you choose those numbers?
Range 30	Q2a	In 30 trials, what might the numbers of reds go from (low to high)?
Range 300	Q2b	In 300 trials, what might the numbers of reds go from (low to high)?
Why on Ranges	Q2c	Why do you think this?
Make Graph	Q3	Make a graph to show what the results might look like for forty trials.

Table 42

Q1 and Q2 related to the aspects of *expecting* and *interpreting*, and will be discussed separately from Q3, which related to the *displaying* aspect.

Results for Q1 & Q2: A comparison of Table 42 above with Table 24 earlier in this chapter will show that PostSurvey (Probability) Q1 matches with PreSurvey Q7, and the tasks have similar names. In the PreSurvey, the context for Q7 was also probability, but a trial consisted of fifty flips of the fair coin, and here in the PostSurvey (Probability) a spinner was used instead of a coin. PostSurvey (Probability) Q2, concerning the expectation on ranges, is

similar to PreSurvey Q2 (which had ranges for 6 trials and then 30 trials). The categories used were the same as those used in the PreSurvey, with no adjustment from PreSurvey Q2 to PostSurvey (Probability) Q1 (since the coin and the spinner represent isomorphic random devices) and with some adjustments in the numbers for range expectations.

Results for Question #1			
Question Number	Coding Level	Description of Category	No. of Students
Q1ai	2	Either gives a range around 25, such as 22-28, or else writes (for example) “Around 25”	10 (34.5%)
	1	Gives 25 blacks as answer	14 (48.3%)
	0	Gives one number other than 25 blacks, such as 23 blacks	5 (17.2%)
Q1aii	3	Uses proportional reasoning with some explicit statement about what else might happen	8 (27.6%)
	2	Uses proportional reasoning (for example: ratio, average, or percent)	17 (58.6%)
	1	Uses additive reasoning , or gives a reasonable response which makes sense but lacks specificity	4 (13.8%)
	0	No reason, a vague reason which makes no sense, or an irrelevant reason.	0 (0.0%)
Q1b	3	[Different or Similar] w/ Explicit mention of a range or spread	15 (51.7%)
	2	[Different or Similar] w/ Some additional information, such as use of ratio, average, percent, or giving specific alternatives for results	10 (34.5%)
	1	[Different or Similar] w/ No additional information provided	2 (6.9%)
	0	No answer, or emphasizes guessing, not knowing, or how results will be the same	2 (6.9%)
Q1c	3	Appropriate choice on Q1ci & Explanation explicitly involves proportional reasoning as well as variation	9 (31.0%)
	2	Appropriate choice on Q1ci & Explanation reflects proportional reasoning or notions of spread	15 (51.7%)
	1	Appropriate choice on Q1ci & Explanation left blank or lacks any specific reasons relating to details of the distribution	3 (10.3%)
	0	Inappropriate choice on Q1ci. W(ide) = Range > 20, N(arrow) =Range < 1, H(igh) = Choices > 24, L(ow) = If Choices < 26	2 (6.9%)

Table 43

Table 43 begins the summary of the categories of responses as well as the numbers of students whose responses were coded at each of the levels.

The presentation of these PostSurvey (Probability) results is streamlined to include the all the sub-parts of a question in a single table, as was done in the PostSurvey (Sampling) whenever questions were similar to PreSurvey questions. More formal comparisons will be made shortly between the result on the PostSurvey (Probability) and other instruments, but for now notice how the responses coded at the lowest level for every question in Table 43 are relatively few in number. Also,the responses at the upper two coding levels make up more than half the total for each question. When guessing for one trial (Q1ai), the most frequent type of response was to just give “25 blacks”, and in reasoning *why* (Q1aii), most students just reasoned proportionally. In comparing trials (Q1b), most students knew that results would likely not be identical, and many gave supporting reasons. The choices for six trials (Q1c) were all appropriate except for 2 which were (W)ide.

The range expectations for Q2 marked the third and final time that this type of question was asked, and the results are in Table 44 on the next page. Similar categories (with appropriate adjustments as necessary) were used in Table 39 for Q3 on the PostSurvey (Sampling), and in Tables 15 and 16 for Q2 on the PreSurvey. Notice how in both Q2a and Q2b on Table 44, more people made appropriate choices than inappropriate choices, an event which had not occurred previously for this type of question.

Results for Question #2			
Question Number	Coding Level	Description of Category	No. of Students
Q2a	1	Appropriate Choice: Min (15 to 20) – Max (30 to 35)	17 (58.6%)
	0	Inappropriate Choice: W(ide) = Range > 20, N(arrow) = Range <14, H(igh) = Min > 20 or Max > 35, L(ow) = Min < 15 or Max < 35	12 (41.4%)
Q2b	1	Appropriate Choice: Min (10 to 18) – Max (32 to 40)	15 (51.7%)
	0	Inappropriate Choice: W(ide) = Range > 30, N(arrow) = Range <14, H(igh) = Min > 18 or Max > 40, L(ow) = Min < 10 or Max < 32	14 (48.3%)
Q2c	3	Explicit mention of increased # of trials leading to increased range, with additional details describing the distribution	5 (17.2%)
	2	Explicit mention of increased # of trials leading to increased range (Note: Could be in the form of mentioning class experience)	16 (55.2%)
	1	Reasoning about one or both ends of the range without explicitly tying the results to the increasing number of trials	6 (20.7%)
	0	No answer or irrelevant answer	2 (6.9%)

Table 44

Application of Q1 & Q2 to Framework (Expecting and Interpreting): As was the case with the PostSurvey (Sampling), so too the responses on the PostSurvey (Probability) questions reflected some of the same dimensions and themes that were highlighted by the parallel questions in the PreSurvey. Therefore, rather than profile each dimension and theme for the aspects of *expecting* and *interpreting* with PostSurvey (Probability) responses, I'll share just a few of the more interesting responses from each question, discussing the responses in the context of the framework.

From Q1a ("One Trial"), here are two responses, the first from DS (who put 28 as her guess for one trial) and the second from EM (who put a range of values):

DS: (28) 28 because there is a 50% chance the spinner will hit either black or white, but it would be very rare that it would hit exactly that many, so I made my guess close to 50% but not perfectly 50%

EM: (22-28) Because there is equal probability to land on black as white, I believe that it will land somewhere close to 50% of the time. From doing the activity in class, I know it won't be exactly 50% but somewhere close.

There is a common theme in both of the above responses that has to do with the expected value of 25. Specifically, both students consider the result of one trial as unlikely to actually be 25. DS says that a result of 25 would be "very rare" and EM sounds certain it won't be 25. In a sense, the "expected value" is no longer *what* is expected, but results should be "close" to that value. EM also picks up on a theme for *why*, as she cites classroom experience in her answer.

On Q1b, comparing a hypothesized second set of results with the first, RL shows *what* he does not expect: The exact same results. Also, DP cites classroom experience as a reason *why* results won't be the same:

RL: The second set of results will likely also be within the same RANGE as the first, but not the exact same results

DP: They will be similar with some variations. In our class experiments, I found when I repeated an experiment you'd often have some new variations pop into the picture but the central probability remains the same

DP mentions the stability of the theoretical probability, but also notes that real experiments result in "some new variations", by which it seems she means different results.

Reasoning on the six trials of Q1c, MA's response reflects a *cause* of variation, while TO's response reflects both a reason *why* for his choice and

also a way of *influencing expectation and variation*:

> MA: (24, 24, 25, 26, 26, 27) I think he will hit 25/50 one time. The rest of the times, he will be close, but not exactly on. Also I think he will be controlling the way he hits the spinner more on the second day, which accounts for no 23 or 28.
> TO: (10, 15, 25, 25, 30, 40) Because due to the data shown in class, the majority of the data will be in the middle but there will be more variety with more data

MA's mention of physical causes of variation helped support the idea I got from other students that a spinner is not considered by some to be a true random generator: That is, spinners can be controlled, in the opinion of MA and others. TO's choices are (W)ide for Q1ci, but her reasoning about most of the data being "in the middle" has some credibility, and she supports her ideas from experiences in class.

Finally, on the expectations for ranges (Q2), notice how DS explicitly relates variation to expectation in her reason *why*, and also incorporates the *influence* of the number of trials:

> DS: (Range for Q2a = 15 to 35, & Range for Q2b = 10 to 40) Because the more trials, the more chance of a larger variation from expected. There are more chances to hit different #s [of] blacks that are close to or that vary from the "expected" theoretical 50%

RL writes about the *influence* of the number of trials on the shape of distribution in his response:

> RL: The more trials run, the more normal the distribution, but the chance of outliers also increases

While Q1 and Q2 on the PostSurvey (Probability) had to do with the aspects of *expecting* and *interpreting*, Q3 asked students to graph 40 trials at the spinner,

and thus addressed the *displaying* aspect.

Results for Q3: This question was similar to Q4 on the PreSurvey, which had students make a graph for 50 trials from the Small Jar in the sampling context. I changed the number of trials to 40 on the PostSurvey (Probability) so that I could avoid having confusion about which "50" meant what if I had instead used 50 trials (with each trial consisting of 50 spins). Coming at the end of the stochastics curriculum in class, I had wondered how many graphs would look like perfectly symmetric bell-shaped distributions centered at 25, and how many showed different degrees of variation.

I used the same categories for Q3 as I had on the similar PreSurvey Q4, and the results are presented below:

Results for Question #3

Coding Level	Description of Category	Number of Students
3	Reasonable variation around the center, vertical scale is also good.	15 (51.7%)
2	Approximate shape for the distribution, centered at or close to 6, but having too much or too little variation. Possibly no attention to vertical scale	10 (34.5%)
1	Attending to every trial (as in a scatterplot or ordinal graph), or graphs in the wrong place w/ no attention to vertical scale	3 (10.3%)
0	No answer or an attempt which is extremely difficult to decipher	1 (3.4%)

Table 45

Most of the class had reasonable graphs (Level 3), although of the 25 graphs coded at the top two levels, 15 of them showed perfect or near-perfect symmetry. I'll show two of the graphs (one each for Level 2 and 3) which had less symmetry than some of the others: A Level 3 pictograph by JL and a Level 2 bar chart by BP (see Figure 12 on the next page). BP's graph

showed some good variation in the frequencies, but having results at both 10 and 45 blacks in 40 trials is too wide.

Before turning to comparisons of the classwide instruments, I also want to look at the graphs in light of the framework, since Q3 represented the third and final opportunity for students' responses to contribute to the dimension of *producing* graphs for the *displaying* aspect. Figure 27 shows two examples of the graphs produced for Q3:

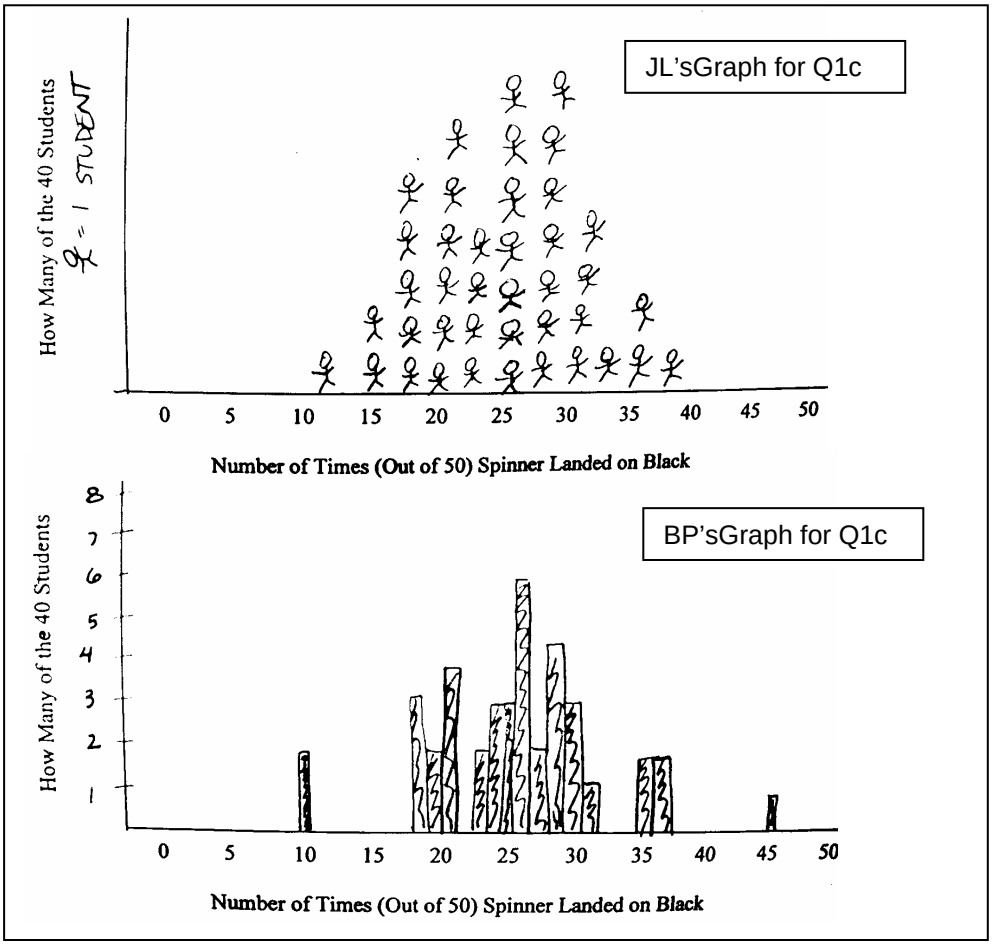


Figure 27

Application of Q3 to Framework (Displaying): Recall that two themes for the *producing* graphs originated in the PreSurvey concerning the technical details and the characteristics of the distribution. Those two themes were later detailed in Table 34 for the PostSurvey (Data & Graphs) Q1c (which had to do with making a graph for average rainfall), and are again used to help look at Q3 for the PostSurvey (Probability). The summary of the Q3 graphs according to the two themes for *producing* graphs is presented in Table 46.

The most common type of graph used was some sort of line plot, usually in form of stacked "x"s, and for line plots the matter of a scale on the vertical axis is moot. On both JL's and BP's graphs (Figure 27), their centers are appropriate as is the general shape. BP's graph, however, has too much spread, since 45 blacks is extremely unlikely to occur in a set of 40 trials with the spinner.

Results for Question #3 (In Relation to Framework)			
Description of Theme			No. of Students
Technical Details	Type of Graph	Bar Chart	7 (24.1%)
		Line Plot (or similar, i.e. - pictograph)	14 (48.3%)
		Bell Curve	4 (13.8%)
		Other (ie, Unconnected Dots)	4 (13.8%)
	Scale on Axis	Adequate	24 (82.8%)
		Inadequate	5 (17.2%)
Characteristics of Distribution	Center	Appropriate	25 (86.2%)
		Inappropriate	4 (13.8%)
	Spread	Appropriate	17 (58.6%)
		Inappropriate	12 (41.4%)
	Shape	Reasonable	22 (75.9%)
		Unreasonable	7 (24.1%)

Table 46

Now that the results have been summarized for the classwide instruments

(the PreSurvey and the PostSurveys), some comparisons will be made for the class on selected tasks .

Comparisons

The purpose in making comparisons with the survey data was to look at any signs suggesting overall trends or shifts in expectation or reasoning which had occurred on a classwide basis. Knowing what shifts might be emerging gave me foresight in subsequently looking at the interview data, and also suggested future directions for more research. The study was designed to look at how individual preservice teachers think about variation, but looking at the survey data with an eye for comparisons for the class as a whole served the research questions by focusing my attention of possible trends or shifts that might arise in the individual cases. In organizing this section on comparisons, I'll detail what I looked at, how I evaluated pairs of questions, and why.

The only candidates for comparison questions were those which had some similarity to the PreSurvey questions, and for which coding levels were used. For example, "One Trial" on the PostSurvey – Sampling (Q2a) was similar to "One Trial" on the PreSurvey (Q1ai), and both were coded along similar levels. The reason for these choices of questions was because I used the coding levels as a way to roughly compare classwide score, and the coding levels used on the PreSurvey have proven useful in other research. Since the categories and coding levels on the PostSurveys were based on or

influenced by those used in the PreSurvey, there is a reasonable expectation that comparisons among the questions using similar or identical coding levels should be useful. Thus, for select questions, the results from earlier tables which were presented will be compared, along with mean class scores.

Again, the point in these comparisons is not to make formal statistical claims, but to get ideas and impressions. Future studies could be designed with larger numbers to do a more comprehensive quantitative analysis, but for the purposes of my research, less formal comparisons suffice. I did include some small-sample t-test results (using unpooled variances) when comparing classwide means, because the smallest *n* for any question was 27 and so such tests were applicable. There could be a question of the assumption of normality for the underlying population distribution, but the robustness of the t-test (combined with the fact that any possible significances that I wondered about happened to align with p-values less than 0.05) made the performance of such tests reasonable. The main criteria in thinking about trends and shifts was not t-tests, but my own impression of what I thought the data might be suggesting: I offer the t-test results as additional support only.

There are two main ways of comparing questions: Within-instrument and across-instrument. By within-instrument (WI) , I mean for example comparing two questions within the PreSurvey, or comparing two questions within the PostSurvey (Sampling). An example of an across-instrument (AI) comparison would be looking at one question on the PreSurvey and its related

question on the PostSurvey (Probability). A secondary way of comparing was by context: There were within-context (WC) and across-context (AC) comparisons. As an example, consider “One Trial” at the Small Jar on PreSurvey Q1ai and “One Trial” at the spinner on the PostSurvey (Probability) Q1ai: The Small Jar is the context of sampling and the spinner is in the context of probability, making the comparison across-contexts.

Since both within- and across-context comparisons apply to both within- and across-instrument questions, there are four types of comparisons that can be made: (WI / WC), (WI / AC), (AI / WC), and (AI / AC). These four types and the specific questions that will be used are summarized in Table 47 on the next page, The organization of this comparisons section is by these types, and the first to be discussed is (WI / AC). Any percentages of students at different code levels are copied from the previous tables where the results were first reported, and specific questions are referred to by the names they were assigned earlier (such as “Compare Trials”, for example). Recall that the PreSurvey was completed by 27 students, the PostSurvey (Sampling) by 30 students, and the PostSurvey (Probability) by 29 students. Missing from the comparisons is the PostSurvey (Data & Graphs), because the questions were different enough from the PreSurvey that comparisons did not seem warranted. Also, recall that coding levels were never used in discussing the PostSurvey (Data & Graphs), and instead a straight application to the framework was made.

Comparisons of Survey Data		
	Within-Instrument (WI)	Across-Instrument (AI)
Within-Context (WC)	1) PreSurvey Q1ci (Six Trials / Sm. Jar) ~ Q2a (Range Six / Sm. Jar) 2) PostSurvey (Sampling) Q2ci (Six Trials / Lg. Jar) ~ Q3a (Range 30 / Lg. Jar) 3) PostSurvey (Sampling) Q3a (Range 30 / Lg. Jar) ~ Q3b (Range 300 / Lg. Jar) 4) PostSurvey (Probability) Q1ci (Six Trials / Spin) ~ Q2a (Range 30 / Spin) 5) PostSurvey (Probability) Q2a (Range 30 / Spin) ~ Q2b (Range 300 / Spin)	1) PreSurvey & PostSurvey (Sampling) [Six Trials] Q1c ~ Q2c 2) PreSurvey & PostSurvey (Sampling) [Several Trials] Q1b ~ Q2b 3) PreSurvey & PostSurvey (Sampling) [Range 30] Q2b ~ Q3a 4) PreSurvey & PostSurvey (Sampling) [Ranges: Reason] Q2c ~ Q3c 5) PreSurvey & PostSurvey (Sampling) [50 Trials] Q3a ~ Q4a 6) PreSurvey & PostSurvey (Sampling) [50 Trials: Reason] Q3b ~ Q4b 7) PreSurvey & PostSurvey (Probability) [One Trial] Q7ai ~ Q1ai 8) PreSurvey & PostSurvey (Probability) [Six Trials] Q7c ~ Q1c 9) PreSurvey & PostSurvey (Probability) [One Trial: Reason] Q7aii ~ Q1aii 10) PreSurvey & PostSurvey (Probability) [Compare Trials] Q7b ~ Q1b
Across-Context (AC)	1) PreSurvey Q1ai (One Trial / Sm. Jar) ~ Q7ai (One Trial / Flips) 2) PreSurvey Q1aii (One Trial Reason /Sm. Jar) ~ Q7c (One Trial Reason /Flips) 3) PreSurvey Q1c (Six Trials / Sm. Jar) ~ Q7c (Six Trials / Flips)	1) PreSurvey & PostSurvey (Probability) [One Trial] Q1ai ~ Q1ai 2) PreSurvey & PostSurvey (Probability) [One Trial: Reason] Q1aii ~ Q1aii 3) PreSurvey & PostSurvey (Probability) [Six Trials] Q1c ~ Q1c 4) PostSurvey (Sampling) & PostSurvey (Probability) [Range 300] Q3b ~ Q2b 5) PreSurvey & PostSurvey (Probability) [Ranges: Reason] Q2c ~ Q2c 6) PreSurvey & PostSurvey (Probability) [Make Graph] Q4 ~ Q3

Table 47

Not every comparison that could be made is represented in Table 47, but just the ones that I was curious about. Details about the comparisons follows next, starting with the (WI / WC) group.

Within-Instrument & Within-Context

The comparisons for this group of questions were not concerning class mean scores along the coding levels, but instead about specific measures. In

comparison (WI / WC) #1, looking at PreSurvey Q1ci ("Six Trials") and Q2a ("Range 6"), I was curious to see if the range that students put on Q2a matched with the range for the six choices that they had already put on Q1ci. In other words, I wondered if it seemed that students were influenced by their previous thinking, or making up a new estimate, and the way I thought of this was in terms of consistency: Were the ranges on the two questions consistent (matching). For the other four comparisons, I wanted to see what percentage of students actually had ranges that did expand when moving from fewer to greater number of trials. The summary for the five comparisons is in Table 48:

Within-Instrument / Within Context			
Comparison		Result	
1)	Range Consistent from PreSurvey Q1ci (Six Trials) to Q2a (Range 6) ?	22% Yes	78% No
2)	Range Expanded from PostSurvey (Sampling) Q2ci (Six Trials) to Q3a (Range 30) ?	87% Yes	13% No
3)	Range Expanded from PostSurvey (Sampling) Q3a (Range 30) to Q3b (Range 300) ?	80% Yes	20% No
4)	Range Expanded from PostSurvey (Probability) Q2ci (Six Trials) to Q3a (Range 30) ?	90% Yes	10% No
5)	Range Expanded from PostSurvey (Probability) Q2ci (Range 30) to Q3b (Range 300) ?	72% Yes	18% No

Table 48

The first comparison above shows how most of the class did not just take their six trials on Q1ci and find a range to put on Q2a, which may be because the questions were on different pages. The result is a reminder how different phrasing of a question can yield different results, and also is a reminder to be watchful for consistency when looking at individual thinking. For the other four comparisons, most of the class consistently put an expanding range for

increasing numbers of trials. This comparison result bolsters the theme that came from the reasoning on ranges, whereby students cited the numbers of trials as influencing the range.

Within-Instrument & Across-Context

Since the PostSurveys were each related to a specific context, such as PostSurvey (Sampling) was related more to sampling, the natural choice for (WI / AC) comparisons was the PreSurvey, which was longer and spanned all the contexts. Table 47 showed how the comparisons in this (WI / AC) group are between the contexts of sampling (at the Small Jar) and probability (with flipping the coin). For these comparisons and all subsequent comparisons in this chapter, the class means are compared. The percentages at each code level are shared in graphical form, the means are given, and the details of the t-test used are provided (the null hypothesis was always one of equal means). For this (WI / AC) group, I had not expected any differences because the questions had all been answered on the same day, prior to instruction. But I wondered if there might be some indication of classwide performance differences because of the different contexts. Figure 28 shows the class results for each pair of questions addressed in the three (WI / AC) comparisons, along with class mean scores (average code level assigned). The PreSurvey and PostSurveys are abbreviated (e.g. PRS for the PreSurvey, POS-S for the PostSurvey - Sampling) Then, Table 49 gives the related details of the t-tests applied, and the comparisons are discussed. The same general format will be used for the comparisons from other groups, such as (AI / WC).

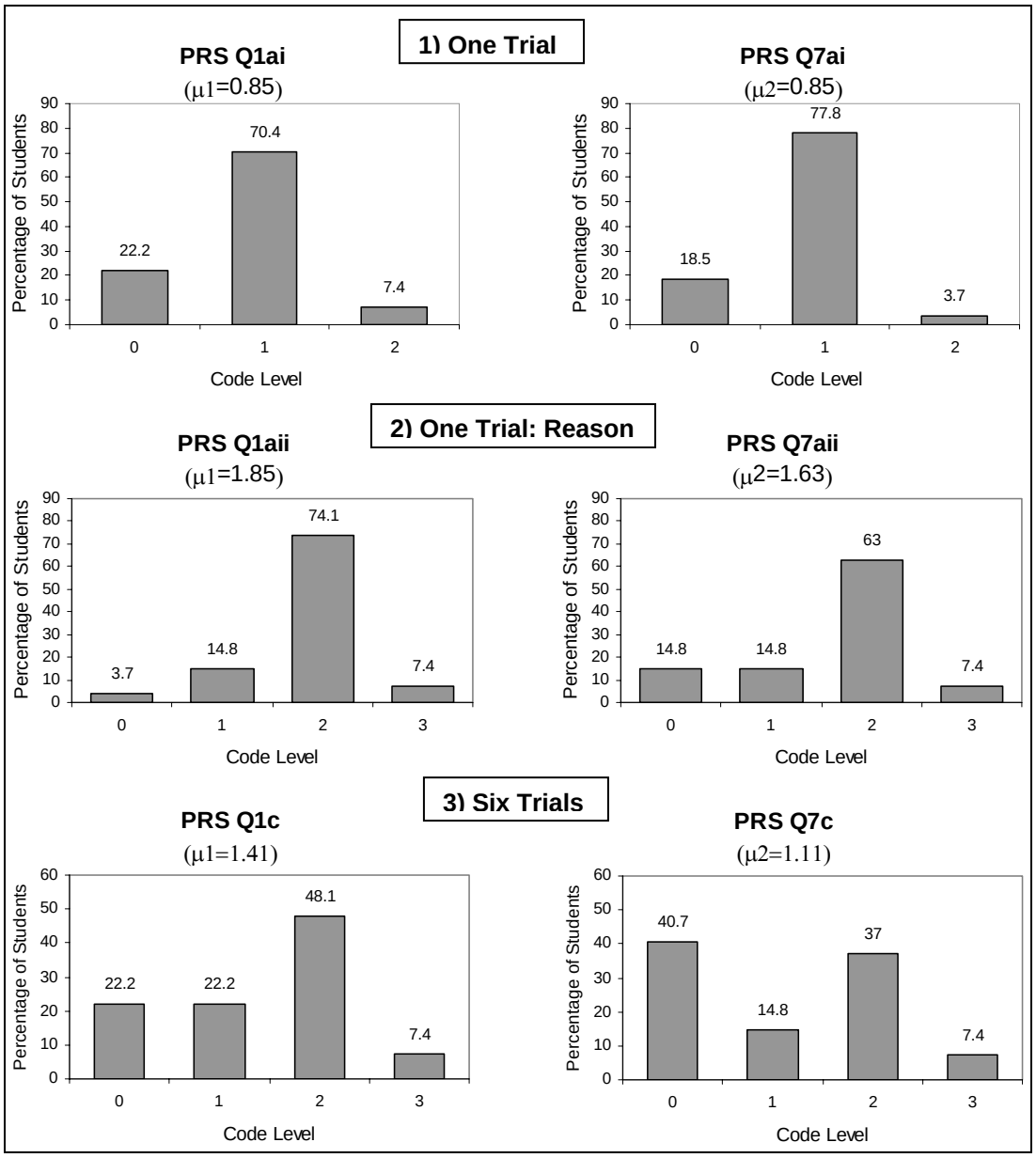


Figure 28

Summary of T-Tests								
	Ha	μ1	μ2	t	p	n1	n2	df
1)	<>	0.85	0.85	0.00	1.00	27	27	50.76
2)	<>	1.85	1.63	1.12	0.27	27	27	47.15
3)	<>	1.41	1.11	1.10	0.28	27	27	51.26

Table 49

The graphs do not suggest to me any significant differences, an opinion

bolstered by the t-tests of unequal means. Although classwide performance seemed stable within the PreSurvey when considering similar questions across the contexts of sampling and probability, in the (AI / AC) comparisons there are some very suggestive differences between sampling and probability when going from the PreSurvey to the PostSurvey-Probability. The apparent lack of differences in the above comparisons becomes more interesting as the later comparisons unfold.

Across-Instrument & Within-Context

There are two sets of comparisons within this group: The first set goes across from the PreSurvey to the PostSurvey (Sampling) in the context of sampling, and the second set goes across from the PreSurvey to the PostSurvey (Probability) in the context of probability. The two sets will be discussed separately.

Across: PreSurvey to PostSurvey (Sampling): My thinking prior to looking at any results was that, on the one hand, classwide results might look different for any related questions, mainly because the PostSurvey (Sampling) was given after activities in class had taken place. On the other hand, the PreSurvey concerned the Small Jar, while the PostSurvey (Sampling) concerned the Large Jar, and so perhaps class results might not be very different – If anything, maybe the Large Jar was even more difficult.

What I noticed after looking at classwide results was that for questions that related more to the dimension of *what* was expected, I did not see

evidence of differences. However, regarding questions having to do with reasoning *why* (which also often allowed for students' responses to reflect some *interpretation* of variation), there did seem to be some evidence of difference. Thus, I'll next show pairs of comparisons corresponding to (AI / WC) #1 & 2, #3 & 4, and #5 & 6 from Table 47. The first comparison in each pair uses questions where students suggest values (such as "Range 30") and the second comparison uses questions where students suggest more reasoning (such as "Ranges: Reasoning").

Figure 29 and Table 50 show graphs and results for the first pair of comparisons, "Six Trials" and "Several Trials".

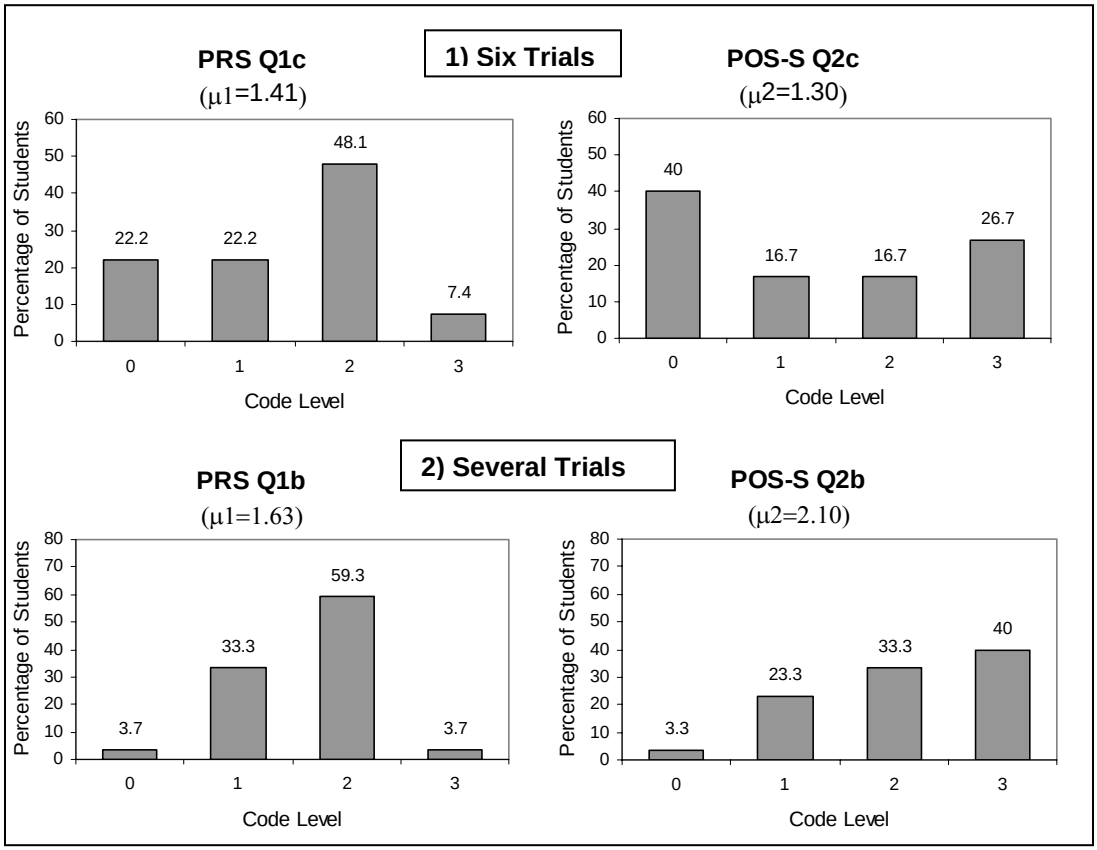


Figure 29

Although “Six Trials” did involve a reasoning component, recall that the criteria of the categories made it hard to have a response coded at a higher level unless the initial choices were appropriate to begin with.

Summary of T-Tests								
	Ha	μ1	μ2	t	p	n1	n2	df
1)	<>	1.41	1.30	0.37	0.71	27	30	53.01
2)	<	1.63	2.10	-2.33	0.012	27	30	52.34

Table 50

The above pair of comparisons is where I started to get the sense that, from the PreSurvey to the PostSurvey (Sampling), there was not as much difference on the questions which involved students picking values as there was on questions where the main focus was on giving reasons. For example, notice above how the reasoning on “Several Trials” (where most people suggested that results would not be identical) seems collectively better on the PostSurvey (Sampling).

This idea of improved reasoning from PreSurvey to PostSurvey (Sampling) gained support from the next pair of comparisons, “Range 30” and “Ranges: Reasons.” On the PreSurvey, the sequence of range expectations went from 6 trials to 30, and on the PostSurvey (Sampling) the sequence went from 30 trials to 300, but I expected similar reasoning in both contexts. The comparison results shown in Figure 30 and Table 51 on the next page do not suggest any difference for the “Range 30” questions: For both PreSurvey Q2b and PostSurvey (Sampling) Q3a, most of the class had inappropriate choices. The figures do not show how 9 choices were coded as (W)ide on the

PreSurvey, and 17 were (W)ide on the PostSurvey (Sampling). So, while numerically the coding levels from PreSurvey to PostSurvey (Sampling) on "Range 30" do not seem different, the subcodes for the categories on those questions showed how more people were (W)ide when thinking of a range for thirty trials at the Large Jar.

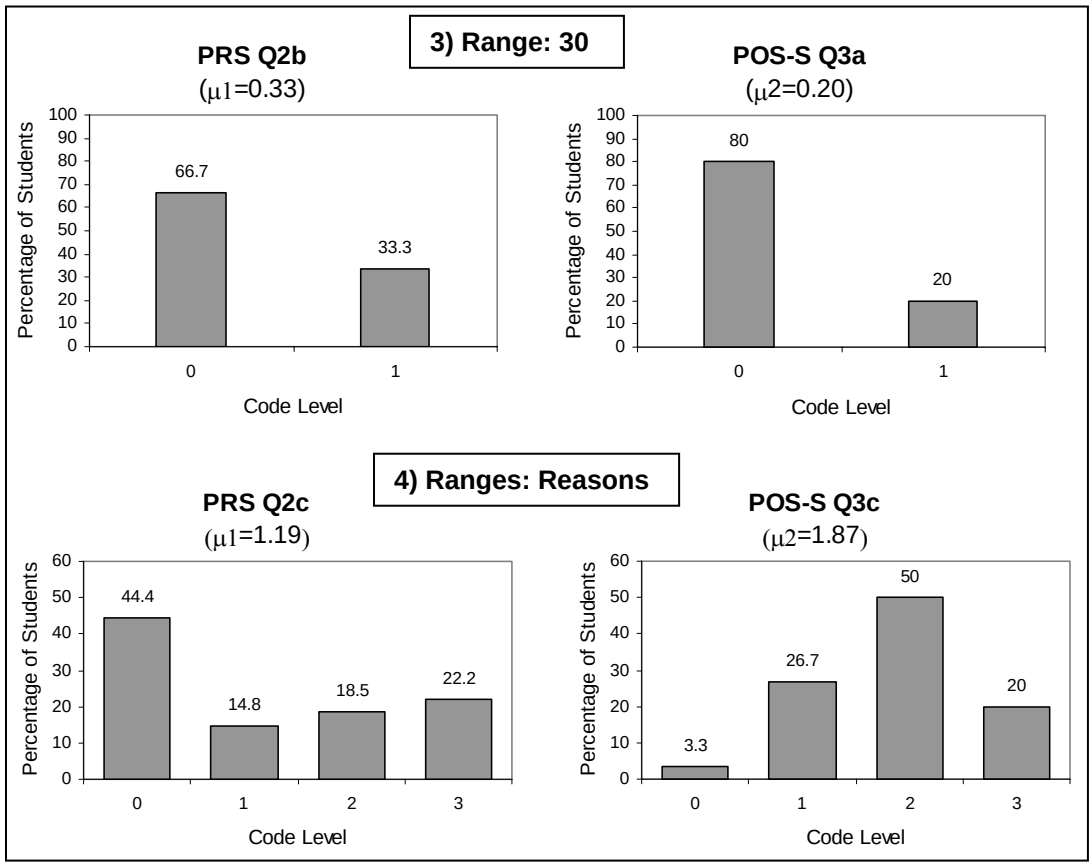


Figure 30

Summary of T-Tests								
	Ha	μ1	μ2	t	p	n1	n2	df
3)	<>	0.33	0.20	1.12	0.27	27	30	51.25
4)	<	1.19	1.87	-2.45	0.009	27	30	42.76

Table 51

For reasoning, there is evidence of classwide improvement, and the final pair of comparisons also supports the trend.

The last pair of PreSurvey to PostSurvey (Sampling) comparisons in this set looked at the actual choices for “Fifty Trials” as well as the reasoning (“Fifty Trials: Reason”). Once again, no apparent differences were seen on the choosing, but there was a shift in the reasoning. The results for the comparisons appear in Figure 31 and Table 52:

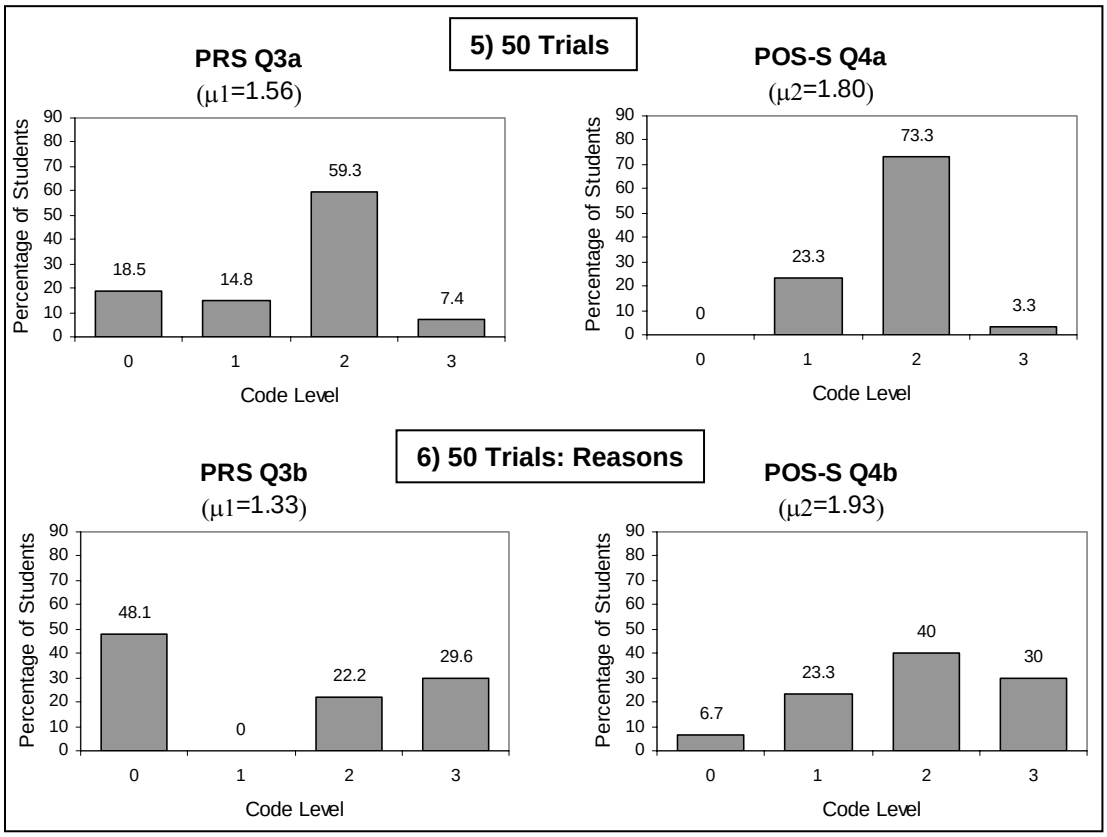


Figure 31

Summary of T-Tests								
	Ha	μ1	μ2	t	p	n1	n2	df
5)	<>	1.56	1.80	-1.27	0.21	27	30	39.16
6)	<	1.33	1.93	-1.94	0.029	27	30	44.61

Table 52

In thinking about why class scores for reasoning might improve while actually making the choices for expectation might not, I thought that perhaps the larger jar of the PostSurvey (Sampling) added a layer of complexity to the questions. For reasoning, however, since the students had been given class opportunities to communicate their thinking, it may be that they gained ideas and the vocabulary to express those ideas from the class experiences. It certainly seemed suggestive that in three separate pairs of comparisons, the actual guessing did not seem to change but the reasoning seemed to improve.

Across: PreSurvey to PostSurvey (Probability): Because this set is within-context, the PreSurvey questions concern the flipping of the coin, while the PostSurvey (Probability) questions have to do with the spinner. Prior to looking at results on comparisons, I wasn't sure if students would react differently to the coin flips versus the spinner, but I thought overall class scores might improve because the PreSurvey was given at the start of the quarter and the PostSurvey (Probability) was the very last survey given. In other words, the longest amount of time was between the PreSurvey and the PostSurvey (Probability).

After looking at classwide results, it seemed that there was overall class improvement for questions that related to the dimension of *what* was expected, as well as improvement for questions having to do with reasoning *why* (which also often allowed for students' responses to reflect

some *interpretation* of variation). I'll next show pairs of comparisons corresponding to (AI / WC) #7 & 8, and #9 & 10 from Table 47.

The first pair focuses on the choosing of values for expectation. That is, the "One Trial" questions had students put what they expected, as did the "Six Trial" questions. "Six Trials" did have a reasoning component, but the coding emphasized the appropriateness of the actual choices. The second pair of comparisons has more to do with reasoning – "One Trial: Reasoning" and "Compare Trials" both focused more on explanations. The comparison results for the first pair are presented in Figure 32 and Table 53.

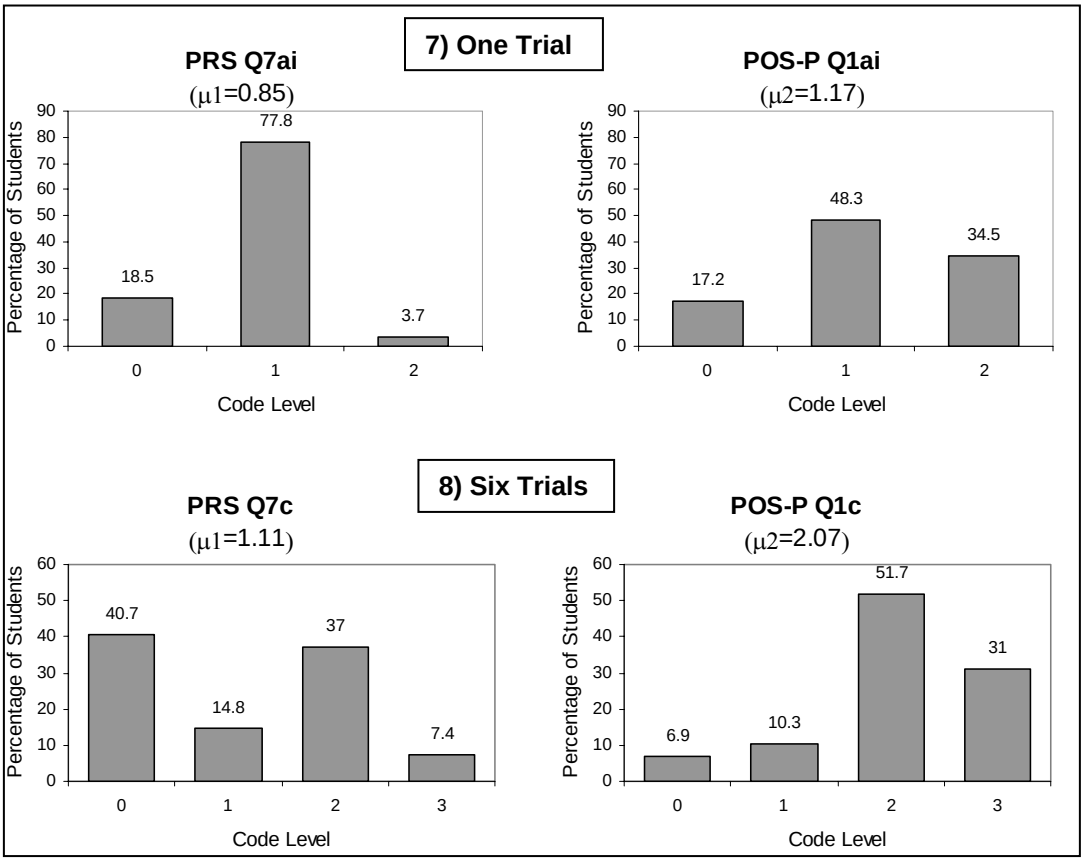


Figure 32

Summary of T-Tests								
	Ha	μ1	μ2	t	p	n1	n2	df
7)	<	0.85	1.17	-2.02	0.024	27	29	48.11
8)	<	1.11	2.07	-3.75	0.0002	27	29	49.86

Table 53

Both pairs of comparisons show a trend toward better overall class results on the PostSurvey (Probability), with a particularly noticeable difference on the "Six Trials" comparison. Again, whether or not any differences may be due to the different probability scenarios (coins versus spinners) or due to the classroom environment is an open question, but there is the impetus to look for shifts in individual thinking going into the interview data.

The results for the second pair of PreSurvey to PostSurvey (Probability) comparisons are given in Figure 33 and Table 54:

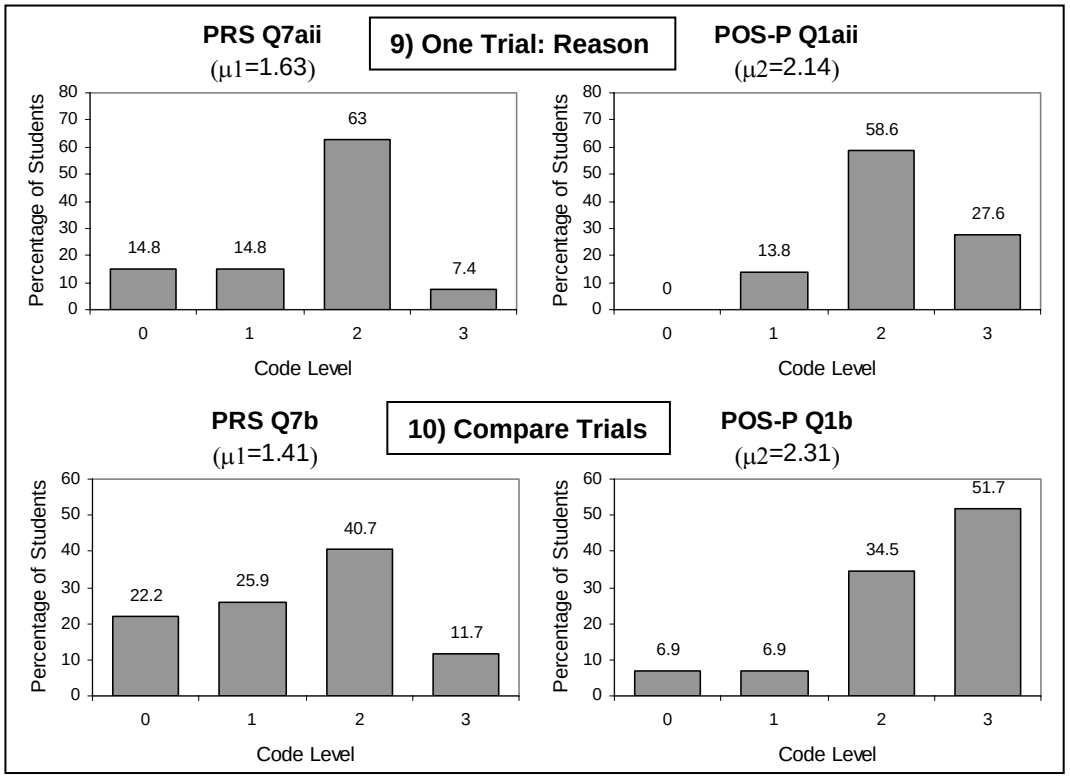


Figure 33

Summary of T-Tests								
	Ha	μ1	μ2	t	p	n1	n2	df
9)	<	1.63	2.14	-2.54	0.007	27	29	48.55
10)	<	1.41	2.31	-3.62	0.0003	27	29	52.67

Table 54

The p-values for this set of comparisons - PreSurvey to PostSurvey (Probability) - are quite low, lending support for the suggestion that the class as a whole got higher scores on the PostSurvey (Probability) as they did on the related questions in the same context on the PreSurvey. Since PreSurvey to PostSurvey (Probability) seemed to show these shifts of classwide improvement in the same context, I was curious to see what results would show on the final group of comparisons, which addressed across-instrument *and* across-context questions.

Across-Instrument & Across-Context

Most of the comparisons in this group are across from the PreSurvey to the PostSurvey (Probability), since that represented the farthest apart two surveys could be chronologically. There was one comparison from PostSurvey (Sampling) to PostSurvey (Probability), for “Ranges: 300”, mainly because those were the only two instruments on which that particular type of question got asked. For all the comparisons in this group, I noticed some shift towards improved classwide scores. The way I’ve organized the presentation of these comparisons in the order the question type was given on the PostSurvey (Probability). For example, on the PostSurvey (Probability), “One Trial” preceded “Six Trials”, which came before “Ranges: 300”, etcetera. The final

comparison for this chapter is about making graphs (PreSurvey Q4 and PostSurvey (Probability) Q3), which spanned the contexts of sampling , probability, and graphs.

One Trial & Six Trials I took the parts of PostSurvey (Probability) Q1 which could be compared to PreSurvey Q1, and those parts included “One Trial” (Q1ai & Q1aii) and “Six Trials” (Q1c). PreSurvey Q1 was in the context of sampling from the Small Jar, and PostSurvey (Probability) was in the context of probability using the spinner. PreSurvey Q1b was phrased in terms of repeating “Several Trials”, while PostSurvey (Probability) Q1b had students “Compare Trials.” Since the categories for coding Q1b were not completely isomorphic, I decided not to compare Q1b from PreSurvey to PostSurvey (Probability). Results for the other three comparisons (Q1ai, Q1aii, and Q1c) are presented in Figure 34 and Table 55 on the next page.

On all three comparisons, the graphs show classwide improvement from the Pre- to the PostSurvey, and the shifts are supported by the results from the t-tests. The p-value for (AI / AC) #3, at 0.004, is particularly low when comparing “Six Trials”, and the graph shows how a much higher percentage of students had responses coded at Level 3 on the PostSurvey (Probability) than on the PreSurvey. The random device on the PreSurvey was a coin, while on the PreSurvey the random device was a spinner, and it may be that students interpret variation differently depending on the device. The change in class performance also could suggest the impact of class experiences.

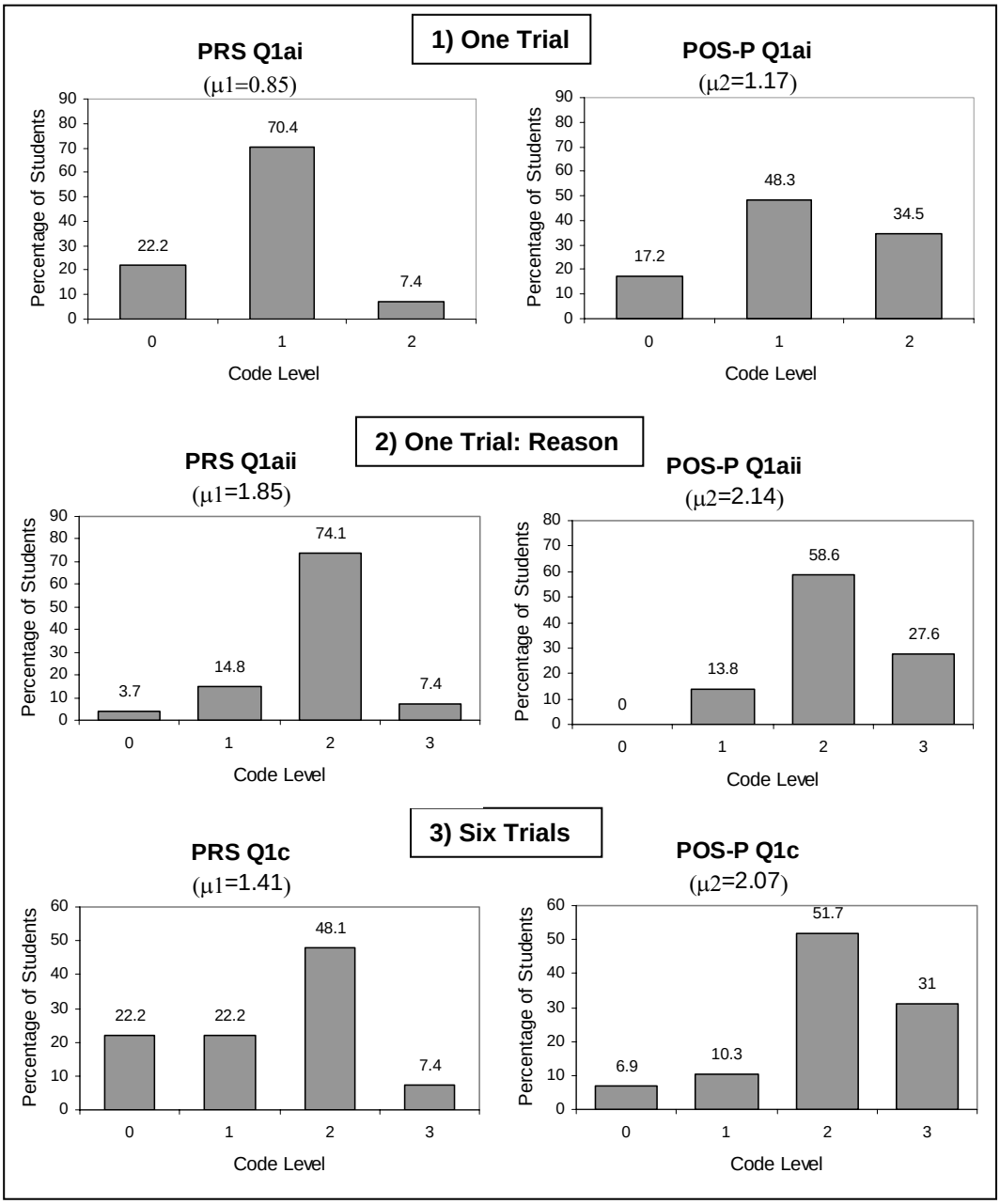


Figure 34

Summary of T-Tests								
	Ha	μ1	μ2	t	p	n1	n2	df
1)	<	0.85	1.17	-1.92	0.03	27	29	51.75
2)	<	1.85	2.14	-1.73	0.045	27	29	53.99
3)	<	1.41	2.07	-2.78	0.004	27	29	52.45

Table 55

Range - 300 & Ranges - Reason: Prior to looking at the results, I had thought that the classwide performance on the two "Range: 300" questions (PostSurvey (Sampling) Q3b and PostSurvey (Probability) Q2b) would be similar, since the PostSurvey (Sampling) and PostSurvey (Probability) instruments were not given out that long apart from each other. The "Ranges: Reason" comparison for PostSurvey (Sampling) to PostSurvey (Probability) did not suggest any differences (and the details are not provided here), but the "Ranges: Reason" comparison from PreSurvey (Q2c) to PostSurvey (Probability) (Q2c) did show an improvement. The two comparisons ("Ranges: 300 and "Ranges: Reason") are shown below:

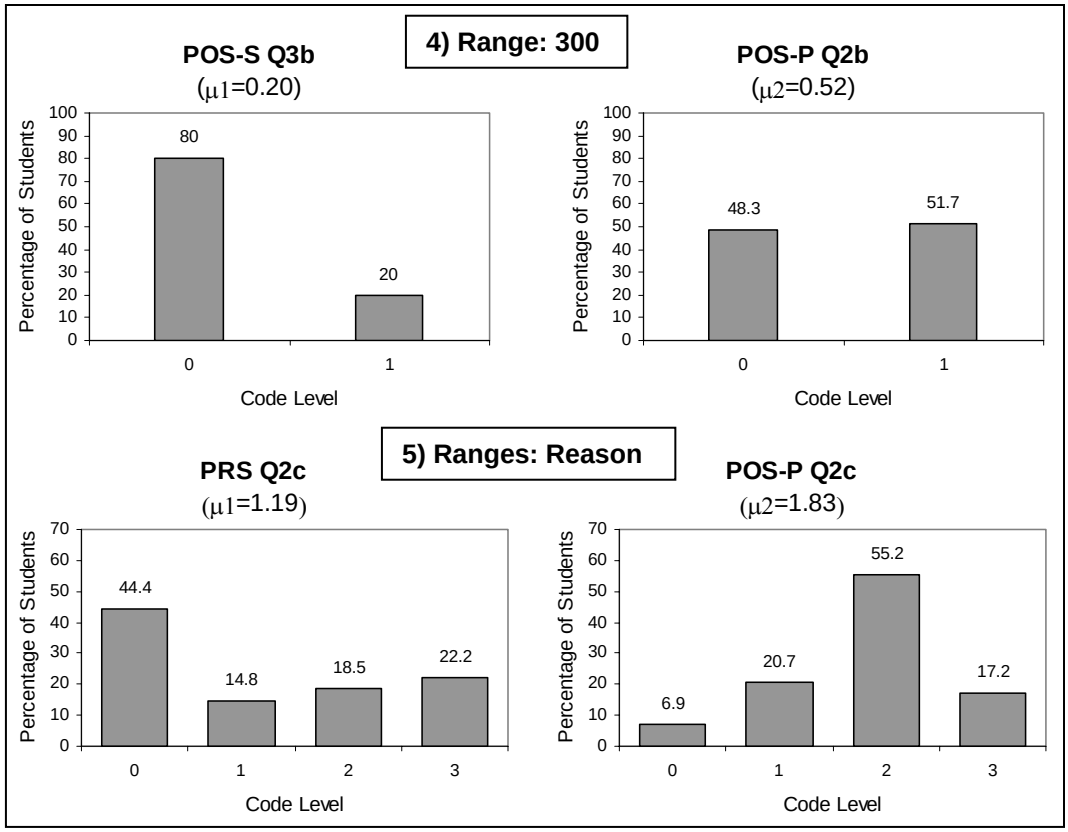


Figure 35

Summary of T-Tests								
	Ha	μ1	μ2	t	p	n1	n2	df
4)	<	0.20	0.52	-2.64	0.005	30	29	53.57
5)	<	1.19	1.83	-2.28	0.014	27	29	44.06

Table 56

Although the p-values in both comparisons above suggest evidence of improvement, visually my attention was drawn more to the comparison for "Ranges: Reason". Seeing how more of the class as a whole went up from Level 0 coding responses was encouragement to think of the class interventions as having had some impact on student thinking.

Make Graph: The final comparison had to do with the *producing* of graphs, involving PreSurvey Q4 (Make Graphs: 50 Trials) and PostSurvey (Probability) Q3 (Make Graphs: 40 Trials). Although the numbers of trials was not the same, the coding categories and levels were the same. In contrast, the PostSurvey (Data & Graphs) question for making a graph was not evaluated using a similar coding scheme: The task for average rainfall was, I thought, very different from the questions in the sampling and probability contexts.

The results of the comparison from the PreSurvey across to the PostSurvey (Probability) are given on the next page in Figure 36 and Table 57. The graphs of classwide results suggest to me a marked difference in classwide performance, an opinion supported by the p-value on the t-test (p = 0.0001). Many more students produced reasonable graphs on the PostSurvey (Probability) than on the PreSurvey, which makes sense given the many opportunities that students had in class to see and make a variety of different

graphs in many different contexts.

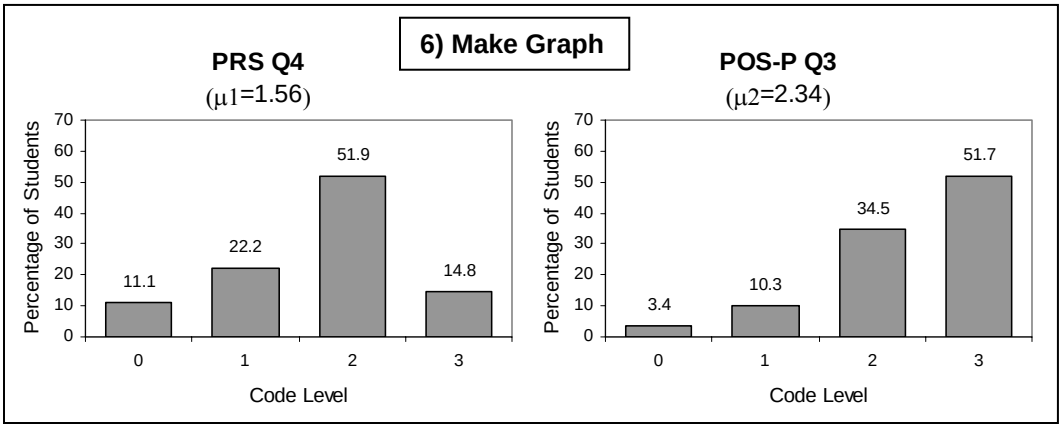


Figure 36

Summary of T-Test								
	Ha	μ1	μ2	t	p	n1	n2	df
6)	<	1.56	2.34	-3.90	0.0001	27	29	53.65

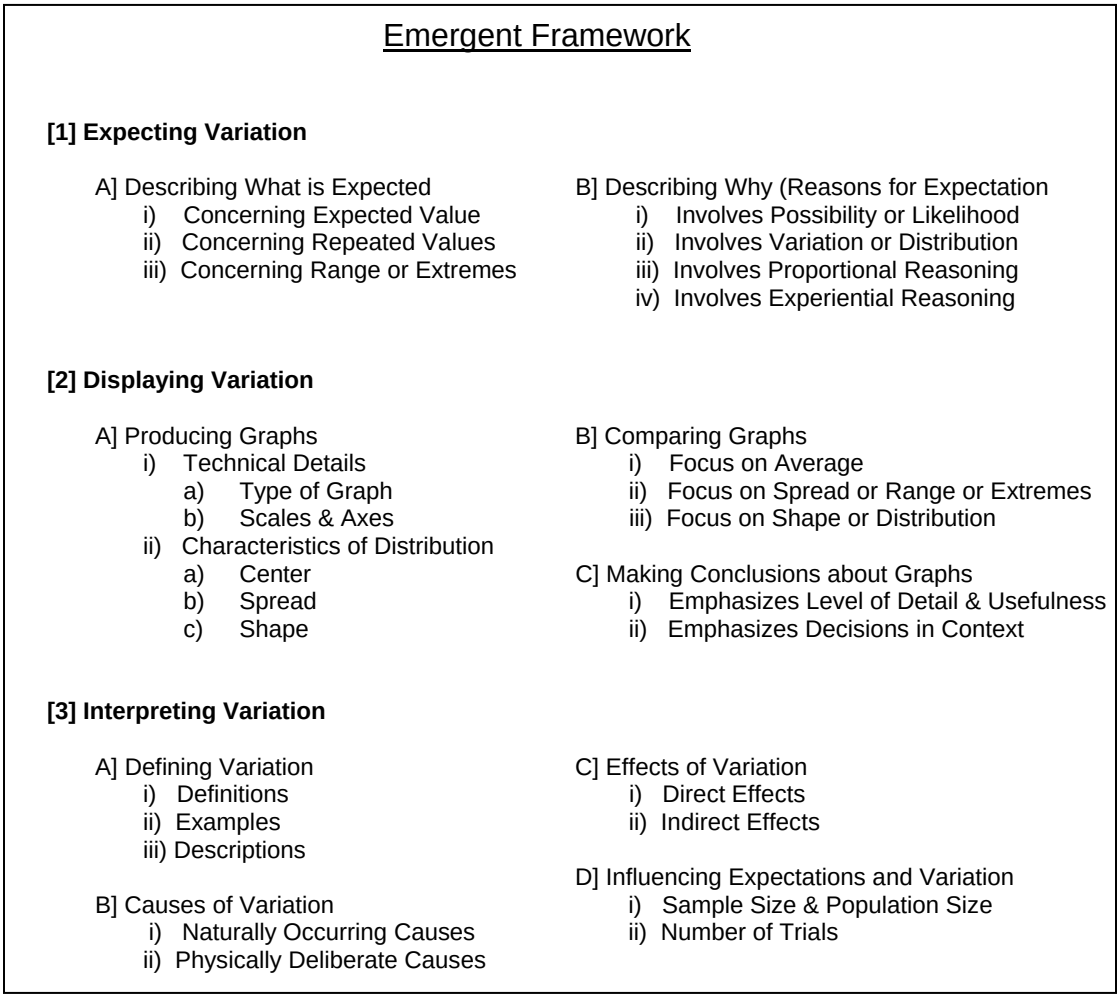
Table 57

The interview data did not cover the *producing* of graphs, but I consider the making of graphs to be connected to reasoning from graphs. Thus, my intention going into the interview analysis was to look for overall improved graph sense, which in the case of this classwide comparison was manifested by improved graph-making.

Conclusion

The results from the classwide data served my research in two different ways. The first way was to give me an overall picture of what kinds of thinking emerged from the class (seen as a case in its own right) on tasks spanning the different contexts for looking at variation, and I was able to get this picture by using the conceptual framework and adding to it. That is, I was able to

use the conceptual framework in its initial form to see that the main aspects and dimensions were useful for looking at student responses. Different themes emerged within the framework, adding depth and providing a richer framework with which to look at the interview data. The revised, emergent framework, with its original aspects and dimensions plus the themes that emerged from looking at the classwide data, is given in Figure 37 below:



Emergent Framework

[1] Expecting Variation

A] Describing What is Expected
- i) Concerning Expected Value
- ii) Concerning Repeated Values
- iii) Concerning Range or Extremes

B] Describing Why (Reasons for Expectation
- i) Involves Possibility or Likelihood
- ii) Involves Variation or Distribution
- iii) Involves Proportional Reasoning
- iv) Involves Experiential Reasoning

[2] Displaying Variation

A] Producing Graphs
- i) Technical Details
  - a) Type of Graph
  - b) Scales & Axes
- ii) Characteristics of Distribution
  - a) Center
  - b) Spread
  - c) Shape

B] Comparing Graphs
- i) Focus on Average
- ii) Focus on Spread or Range or Extremes
- iii) Focus on Shape or Distribution

C] Making Conclusions about Graphs
- i) Emphasizes Level of Detail & Usefulness
- ii) Emphasizes Decisions in Context

[3] Interpreting Variation

A] Defining Variation
- i) Definitions
- ii) Examples
- iii) Descriptions

B] Causes of Variation
- i) Naturally Occurring Causes
- ii) Physically Deliberate Causes

C] Effects of Variation
- i) Direct Effects
- ii) Indirect Effects

D] Influencing Expectations and Variation
- i) Sample Size & Population Size
- ii) Number of Trials

Figure 37

The second way that the classwide data helped my research was by calling attention to important trends in thinking that were useful when looking at the

interview data. That is, by using a coding scheme that had already been established as useful in looking at middle and high school students' thinking, I was able to look at some comparisons involving the performance of the class as a whole on some of the questions from the Pre to the PostSurveys. In particular, I noted how reasoning seemed to improve from the PreSurvey to the PostSurvey (Sampling), and reasoning as well as making predictions and graphs seemed to improve for the class from the PreSurvey to the PostSurvey (Probability). Thus, I left the analysis of the classwide data thinking that it was a reasonable assumption to expect some degree of differences in the interview data from before the class interventions to after (that is, from the PreInterview to the PostInterview).

In moving now to the PreInterview and then the PostInterview data, the revised framework will be used to look at not all of the class, but only the six students that were chosen as my individual cases. These results and analyses are given in the next chapter.