Transition to PostInterview

The preceding sections of this chapter have allowed the conceptual framework to develop by considering responses of all six cases to all the questions on the PreInterview. In the subsequent sections, I'll use the framework as a lens to examine the case study PostInterview data, emphasizing the individual thinking of each case. Each case will be organized according to the aspects in the framework. That is, first the aspect of expecting will be discussed with respect to the case, followed by the aspects of displaying and interpreting.

Rather than use each question on the PostInterview, as was done for the PreInterview, a subset of six questions was selected from among all those asked on the PostInterview, and this subset became the common basis for looking at each case. The specific questions, described in the next section, were chosen for three reasons. First, the cases' collective responses on these questions spanned the same aspects and dimensions that were brought out in the PreInterview. This means that both the PreInterview and PostInterview responses can be used to support the categories in the framework. Second, the six questions span all the contexts that were of relevance throughout the research – namely, sampling, data and graphs, and probability. Third, three of the questions are similar to questions which were asked earlier in either the PreSurvey or the PreInterview. This similarity makes it useful to compare responses across different phases of the study. The other three questions are unique to the PostInterview, and offer new opportunities for the cases to reveal

thinking about variation. Additional questions from the PostInterview may be brought into the discussion if they bring some special relevance to a case.

PostInterview Questions

Although the entire PostInterview is listed in Appendix B, the six questions that form the common subset for looking at all the cases are also given in this section and briefly discussed before turning to the case studies.

Question 1

The structure of Question 1 (see Figure 43) is similar to Q1 in the PreInterview and also Q1 in PreSurvey. However, the big change is that in the PostInterview, the jar is no longer a 60 Red and 40 Yellow mix, but a 600 Red and 400 Yellow mix.

PostInterview Q1
(Similar to PreSurvey Q1 & PreInterview Q1)

[1] Suppose there is a large container with 1000 pieces of candy in it. 600 are Red, 400 are Yellow. The candies are all mixed up in the container. You reach in and pull out a handful of 100 candies at random.

(a) "One Trial"
How many red candies do you think you will get?
(b) "Repeated Trials"
Suppose you do this several times (each time returning the previous handful of 100 candies and remixing the container). Do you think this many reds would come out every time? Why do you think this?
(c) "Six Trials"
Suppose six classmates do this experiment (each time returning the previous handful of 100 candies and remixing the container). Write down the number of reds that you think each classmate obtained. Why did you choose those numbers?

Figure 43

In other words, the PreSurvey and PreInterview questions were based on what I called a "Small Jar", while in the PostInterview the similar question was based on a "Large Jar." Also, in the Small Jar only 10 of the 100 total

candies were drawn in a trial, but in the Large Jar a handful of 100 out of the 1000 total was pulled in each trial. The context for this question was sampling, and elicited responses from the six cases in both the expecting and interpreting aspects.

Question 5

This distinction between Small and Large Jars was important for Question 5 ("Small & Large"), because this was the only question in this research that directly appealed to a change in population size while keeping the sample-to-population ratio fixed. The statement of the question is in Figure 44.

PostInterview Q5

On a day of planned absence from school, Keith left these instructions for his two classes:

- He told the forty students in "Class A" to go to the SMALL container (100 Candies = 60 Red and 40 Yellow). They were each supposed to draw small handfuls of 10 candies (with replacement after each draw of 10).
- He told the forty students in "Class B" to go to the LARGE container (1000 Candies = 600 Red and 400 Yellow). They were each supposed to draw small handfuls of 100 candies (with replacement after each draw of 100).

When Keith came back the next day, he saw these graphs and sets of data for the two classes (See Figure 45 on next page). Keith suspects that one of the two classes just made up the data and didn't really carry out the experiment. What do you think? That is, based on the two graphs shown above, do think one graph is likelier than the other to reflect made-up data ?

Figure 44

The bar graph for the 40 trials from the Small Jar was in fact created from real data, using the Fathom software to simulate the situation. The bar graph for the 40 trials from the Large Jar is fabricated, and highly unlikely to have resulted from actual data. The graphs are shown in Figure 45

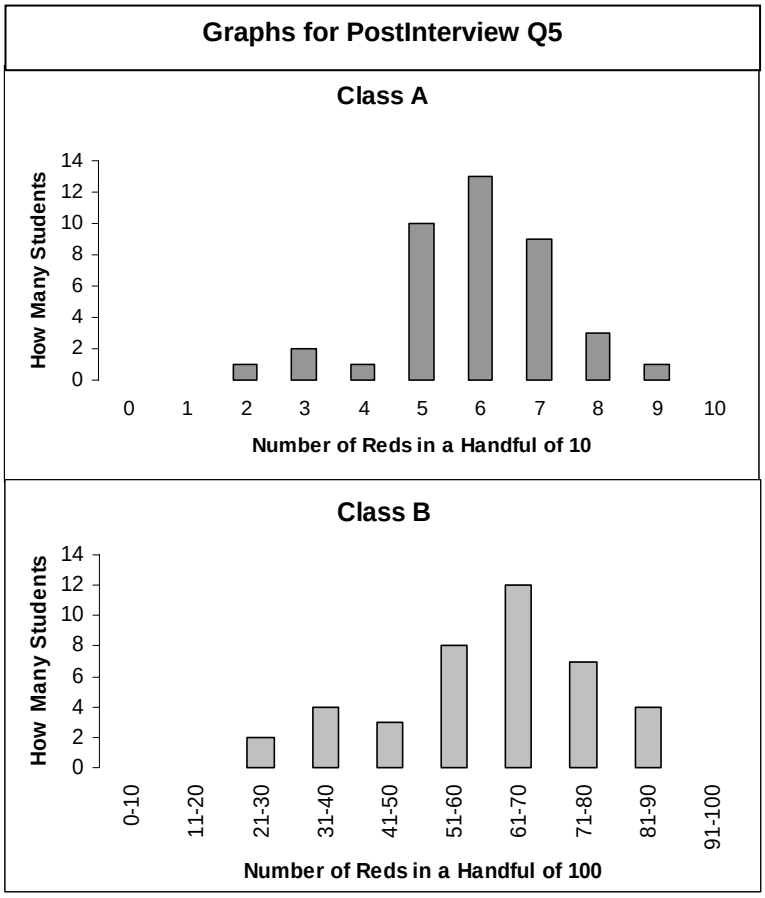


Figure 45

The important thing about the two graphs is that in terms of shape, they were extremely similar. The contexts included graphs, since two graphs were provided, but again the main context was sampling. All three aspects were touched upon by this question.

Question 8

For Q8 (“35 Muffins”), the context was data and graphs. The data comprised weights in gram (rounded to the nearest half-gram) for 35 different muffins bought at different times during the week from the same bakery. In addition to listing the data in a table, and showing it in a boxplot and

histogram, a box of summary statistics (mean, median, mode, maximum and minimum) was given, as shown in Figure 46.

PostInterview Q8

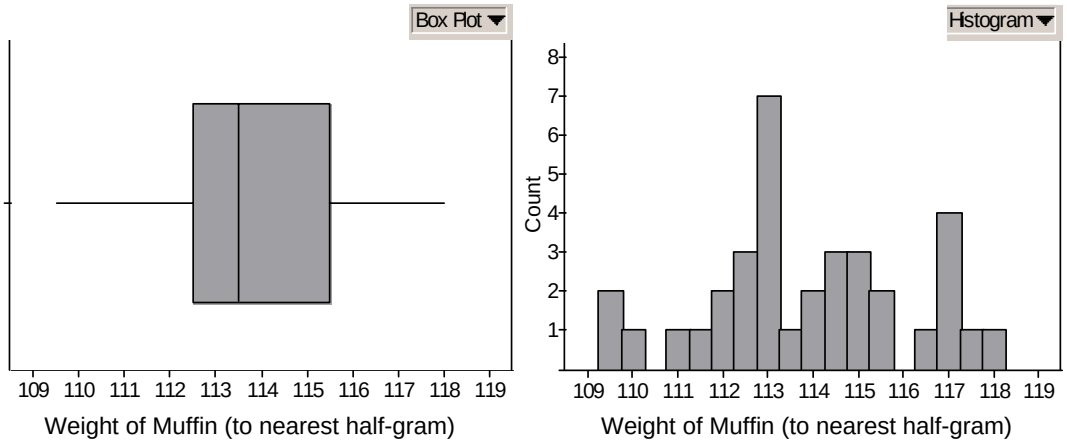
A different class of 35 students wanted to know how much a "Grande Muffin" from the West End Bakery weighed.. They decided that each student should buy one "Grande Muffin" at different times during the week, and weigh the muffins before eating them.

They recorded their 35 muffin weights (rounded to the nearest half-gram, see below), and summarized the results with a boxplot and a histogram:

Data: (Weight of Muffin)

109.5	109.5	110.0	111.0	111.5
112.0	112.0	112.5	112.5	112.5
113.0	113.0	113.0	113.0	113.0
113.0	113.0	113.5	114.0	114.0
114.5	114.5	114.5	114.5	115.0
115.0	115.0	115.5	115.5	116.5
117.0	117.0	117.0	117.5	118.0

Mean = 113.79 g
Mode = 113.00 g
Median = 113.50 g
Minimum = 109.50 g
Maximum = 118.00 g



a) Suppose you bought a "Grande Muffin" from the West End Bakery. How much do you think your muffin would weigh? Explain your reasoning.

b) What are some similarities and differences in the way these two graphs present the data?

c) Do you think one graph tells you more about the variation in the data than the other graph? Explain your thinking.

Figure 46

This question invites not just a point estimate, but an opportunity to discuss likely ranges, given the variation shown in the data set. Also, they are asked

explicitly for some similarities and differences in the way the two graphs show the data, and this was done because I was interested if their comments would reflect the way in which different graph types reveal or mask variation. The main aspect from this question was that of displaying variation, with a small number of responses showing the aspect of expecting as well.

Question 9

The similarities between Q9 ("Two Bakeries") on the PostInterview and Q8 ("MAX Wait-Time") on the PreInterview lay in the fact that both questions gave two sets of data, each with different amounts of variation. For the MAX data, the means and medians were identical, but on the PostInterview there are differences in the averages (see Figure 47).

PostInterview Q9
(Similar to PreInterview Q8)

The MathTeam must choose a new bakery to supply them with muffins. The MathTeam wants muffins that are usually at least 110 grams in weight.

So, the MathTeam samples twelve muffins from the West End Bakery and twelve muffins from the East Side Bakery. The weight of the muffins from the different bakeries were as follows:

West:	93	102	102	108	Mean	=	110.25
	111	111	114	114	Median	=	112.50
	114	117	117	120	Mode	=	114.00
East :	88	91	91	95	Mean	=	110.75
	95	99	104	104	Median	=	101.50
	136	142	142	142	Mode	=	142.00

i) Do you think that one bakery would better meet the needs of the MathTeam over the other bakery? Explain your thinking.

ii) Is one graph or the other more helpful in showing the differences between the two bakeries (see Figure 48) ? Explain your thinking.

Figure 47

A Boxplot and Dotplot were used to portray the data (see Figure 48).

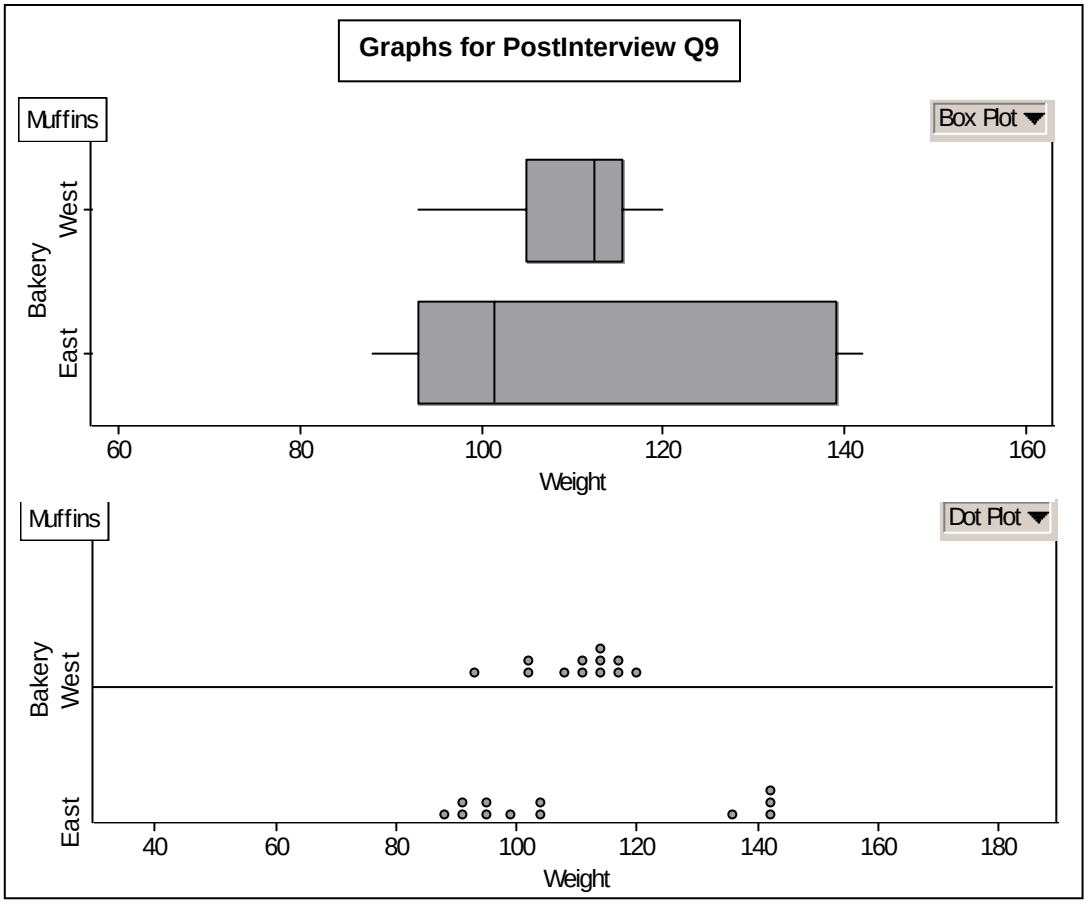


Figure 48

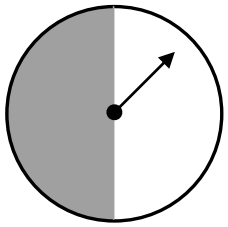
Students are asked how the bakeries compare to one another in terms of the weights of the muffins they produce, with the intention of eliciting comments on the noticeable differences in variation shown in the two graph types, boxplots and dotplots. The students are also asked explicitly if one graph or the other is more helpful in showing the differences, and this was done so I could get a sense of their preference for graph type with respect to displaying variation. Again, this question was in the context of data and graphs, and all three aspects were informed by responses to this question.

Question 10

Probability was the context for Q10 in the PostInterview (see Figure 49), which was similarly structured to Q11 in the PreInterview. The PreInterview scenario used 60 tosses of a fair die, whereas in the PostInterview a single trial involved 50 spins of a half-white and half-black spinner. PostInterview Q10 also mirrored PreSurvey Q7, which used 50 flips of a fair coin.

PostInterview Q10
(Similar to PreSurvey Q7 & PreInterview Q11)

Consider the spinner on the right:



a) "One Trial"
Matt is curious to see how often the spinner lands on black, so he spins it 50 times. How many times (out of 50 tries) do you think the arrow might land black? Why do you think this?

b) "Compare Trials"
After Matt's first set of 50 spins, he decides to do a second set of 50 spins. How do you think his results on the second set of 50 spins will compare with the results of his first set?

c) "Six Trials"
Matt actually has a lot of time on his hands, so the next day he does 6 sets of 50 spins. Write a list that would describe what you think might happen for the number of spins out of 50 the spinner would land on the shaded part in each of the 6 sets of 50 spins. Why did you choose those numbers?

Figure 49

The "One Trial" subquestion (Q10a) was followed by a subquestion comparing results on a second trial with the first trial (Q10b), followed by a "Six Trials" situation (Q10c). The aspects of expecting and interpreting variation were informed by the cases' responses to Q10.

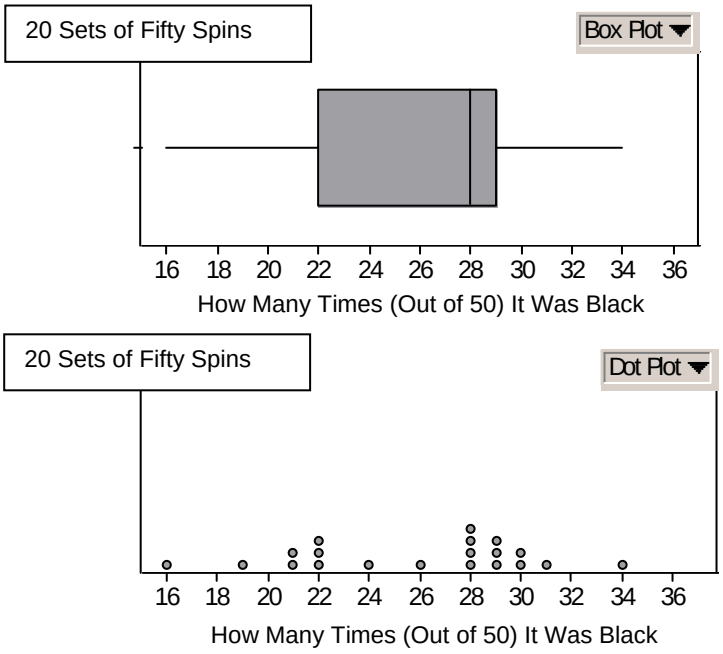
Question 12

Lastly, involving contexts of probability and graphs, another question

unique to the PostInterview was Q12 (“Compare Comments”). This question shows data from 20 trials of the spinner, focusing on the number of times the spinner shows black in each of the 20 sets of 50 spins. The data is presented in a boxplot as well as a dotplot, and shows actual results from a Fathom simulation (see Figure 50).

PostInterview Q12

Twenty students lined up at the spinner. Each student spun it 50 times and wrote down how many times (out of 50 tries) the arrow landed on the shaded part. At the end of class, they decided to make graphs of their data. Their boxplot and dot plot looked like this:



* Keith argued that something was wrong with the experiment because no one got exactly 25 out of 50 spins landing on black.
* Karen argued that, since the mode (and median) was 28 out of 50 spins landing on black, something was wrong with the spinner.
* Jeanette argued that the maximum of 34 out of 50 spins landing on black would not have happened unless something was wrong with the spinner.
* Marjorie argued there was nothing wrong with the spinner, since she had expected the results to look like this.

(a) What do you think about each person’s argument?

(b) Is there a different argument that you would make? Explain.

Figure 50

While the range (from 16 black to 34 black) is unremarkable for an actual experiment, it is interesting that none of the 20 trials resulted in a single occurrence of the expected value of 25 blacks out of 50 spins (corresponding the way that half the spinner is black). The scenario provides four sample comments from among the people who supposedly carried out the experiment: I then asked each of my six cases to respond to or compare the comments, or to make their own argument about the results. This was the only time in the research when a format was used that provided specific details for the cases to consider, and this was done because sometimes a subject can speak about a topic in different way if they feel they are reacting to what someone else has said.

The six questions, taken together, form the common subset which illustrated whatever aspects and dimensions came from the following six cases. These cases are presented next, organized by the aspects of the framework.

The Case of DS

A brief discussion of DS is given before focusing on the aspects brought out by her responses on the PostInterview, and this is done to situate her responses within the larger context of the tasks and instruments which had preceded the PostInterview. Further discussion takes place at the end of the case study.

DS was a very energetic individual who readily expressed opinions and

thoughts on all the questions. She had taken Math 211 the previous quarter with Steve, and had also taken a prior course in probability and statistics at another college, saying she "loved it." On the PreSurvey, for her initial definition of variation she had said that variation meant "changes over time," and cited her mood as an example of something that varies.

There were also several indicators given in subsequent questions on the PreSurvey showing that DS had a good grasp for the basic ideas involved in probability and statistics. For example, her "Pick Six" (PreSurvey Q1c) on the sampling context was reasonable, and in moving from six trials to thirty trials, her conjectured ranges expanded slightly (PreSurvey Q2 – "Ranges"). In comparing graphs (PreSurvey Q5) she used means, and she was consistent in reasoning when going from the sampling context to the probability context on the PreSurvey.

During the PreInterview, DS mentioned at several instances the connection between the number of trials and the amount of variation, and in these instances she was primarily thinking of variation in terms of an increased range. This helps explains why, in PreInterview Q3 ("Fake: 30"), she was willing to think of the graph for 30 trials from the Small Jar as realistic. Thirty trials were not very many, and so it seemed plausible that "they just didn't have a lot of variation" and hence only got either 5, 6, or 7 reds in their handfuls. Another theme repeated by DS throughout the PreInterview was that results won't always be "perfect," which was her way of expressing what the laws of probability predict in given situations. She often spoke of

probability in terms of "odds" (suggesting for instance that the expected value is what the odds would predict), but repeated the idea that she could get more or less than her prediction. She was generally confident in most of her remarks, and on the occasions when she would reconsider a previous opinion and change her mind, her latest conclusions were again expressed with confidence.

Expecting Variation

Of the six PostInterview questions looked at for the case studies, five of them drew

responses from DS informing this aspect. Q9 ("Two Bakeries") was only a part of the displaying aspect. First the dimension of *what* was expected will be discussed, followed by the dimension of *why*, which suggests reasons for expectations.

What was Expected: Concerning the expected value, DS was very clear that results should be close that value. For example, in Q1, drawing 100 candies the Large Jar gives an expected value of 60 red. In Q10 and Q12, a set of 50 spins gives an expected value of 25 black. In each of Q1, Q10, and Q12, DS mentions results being close to, or around, the expected value:

DS: [Q1c] You'll probably stay close to the average {60 Reds}.
DS: [Q10] There's a … 50% would be 25, and … Or maybe 23. So, somewhere close to 50 {%}, I mean, to 25, but not…Like, just close to 50%, but not exactly!
DS: [Q12] I think that you know, you're going to be close to the half number, 25.

The above quotes are not the only comments by DS within each question that

get at the theme of being close to the expected value, but they are representative of her thoughts. The style in presenting the case studies will be modeled on the premise that reporting sheer numbers of comments are not as important at this point as identifying ways of thinking within each aspect. In the next chapter, a comparison is drawn between the cases which does rely in part on the frequency of comments with respect to the different aspects. Note that even though DS suggests results should be close to the expected value, she also views this value as unlikely in Q10, as she remarks that "I think it'd be rare that we'd get exactly 25 on our first spin." Also, in considering the two graphs in Q5 ("Small & Large"), DS makes it clear that she expects less data when moving farther away from the expected value:

> DS: [Q5] Because I think, um, you know, {She's pointing to Class B} as you get farther from average number of reds – the 60 – then there are going to be fewer, fewer people drawing that number.

Finally, for DS the notion of average and expected value has some congruence in the sampling and probability contexts. It was therefore no surprise that in the "Two Bakeries" scenario of Q8, DS had this to say about expectations for the weight of

her muffin:

> DC: [Q8] Suppose you buy a Grande Muffin from the West End Bakery, how much would you expect your muffin to weigh?
> DS: About 113 and a half grams {Laughs}

Of course, the boxplot as well as the summary statistics gave the median

weight of the other 35 muffins as 113.50 g, and so it made sense that DS would expect a weight close to that central value.

Another theme about *what* was expected had to do with if or how much values might repeat in a set of results. In Q10b ("Compare Trials"), DS is explicit that she expects a second trial to have different results from the first trial:

> DC: [Q10b] How does his result on the second set compare with the results on his first set?
> DS: I think it'd be close to it, but different. So, maybe if he got 28 the first time, he'd get 24 the second time, or 23…

Consistent with expectation of results that don't repeat, when it was time for DS to "Pick Six" on Q1c and Q10c, none her six choices had repeated values. She put "64, 58, 61, 56, 60, 62" for Q1c, and "28, 23, 25, 29, 21, 26" for Q10c. In each instance, the expected value does appear, but we also get a sense of what being close really means for DS. It is also understood that she doesn't expect values to repeat, at least not for six trials.

Concerning the range, or extreme values, DS commented in both Q1 and Q10 about how results should be within some kind of range, and this idea fits with her earlier remarks about results being close to the expected value but different from one another. She has in mind some range for each situation within which she is comfortable seeing the results fall, such as the 23 – 28 that she mentioned earlier in Q10b. For doing trials from the Large Jar in Q1, she also specifies a range:

> DS: [Q1b] You'll have more in the center, around the 60 number. {Lots of gesturing to make her point}

DC: Yeah, I'm going to try to figure out this "around 60", what you mean by that, so…
DS: Like… Probably 55 – 65, I mean, yeah…

Notice how her "Pick Six: in Q1c ranges from 56 to 64, falling within the range she suggested. Equating a range with showing variation, she expressly states an expectation of variation in the Q10: "Yeah, and then I have a few scattered close to 25, but not 25, so that… 'Cause there's gonna be variation." However, DS still attends to extreme values, being skeptical if there are too many, as she did in Q5 ("Small & Large"). Earlier DS is quoted in reference to Q5 about how she expects less data when moving farther away from the expected value, and when she looked at the data in the minimum and maximum bins for Class B drawing from the Large Jar, she adjudged that graph as having too many extremes, suggesting that she would have expected less.

In summary, what DS expects her results to look like is that they should be close to the average or expected value, but not necessarily the same results are expected on repeated trials. However, results should stay with a range, with too many extreme values being suspect. With this picture in mind, it is not surprising that her responses to Q1a and Q10a ("One Trial") in the sampling and probability contexts were not the expected value. She put down "64" for one trial from the Large Jar and "28" for one trial at the spinner.

However, it is in a sense a very natural thing for a person to list the expected value when asked for expectations on one trial. In fact, the expected value is exactly what DS gave on the similar questions in the PreSurvey and

PreInterview: She put “6” for the one trial at the Small Jar (PreSurvey Q1a and PreInterview Q1a), and also put “25” for one trial of fifty flips of the fair coin (PreSurvey Q7a). Yet by the time of the PostInterview, her responses to expectations for one trial were no longer to simply state the expected value.

Why (Reasons for Expectations): The first theme in this dimension has to do with likelihoods and possibilities – That is, how likely or unlikely an event is to occur, or the chances or possibilities of a certain result. As mentioned earlier in this chapter, there can be at times a certain overlap for some comments which address both *what* is expected and *why*. DS provides an example of this overlap in her response to Q5. When she was looking at Class B and reflecting that it showed too many extremes, she said that the results were “unlikely”, and thus we see that DS does not expect that many extremes (showing *what* she does or does not expect) and also why (because they are unlikely). However, extremes are still possible, as she shows in Q12 (“Compare Comments”), when DS responds to Jeanette’s claim that 34 blacks out of 50 spins shows evidence of an unfair spinner: “Well, the 34 is 9 away from 25, so, you know, only one person got that, which definitely could happen.” Also, recall how earlier it was shared that DS mentioned in Q10 that “it’d be rare” to get the expected value of 25 black on her first trial at the spinner, and now in Q12 more of a rationale is given:

> DS: [Q12] There’s no guarantee that there’s.. That you’re going to get exactly 25 out of 50. Because, the spinner is… Um, randomly landing.

Notice how, in DS's response, an effect of variation is also seen – The random nature of the spinner means there is "no guarantee" of getting the expected value, which earlier in the PreInterview had the flavor of being the "perfect" result for DS.

DS does show a facile use of proportional reasoning, which she demonstrated in Q1 and Q10 by referring back to the original ratio of candies in the Large Jar, or the ratio of white to black on the spinner. It did not seem that DS overly relied on this form of reasoning, which can tend to fixate some subjects' attention on the expected value. DS knew how to reason proportionally, and she used this knowledge as a point of reference. For example, in Q10 she noted that "You have 50-50 blacks and whites. So, out of 50 spins, you have half chance of getting blacks and half white." Occasionally she would use proportional reasoning as a way of justifying or explaining why results should be close to the expected value.

Lastly, some of DS's reasons explicitly relate to variation or the distribution of the data. For example, in justifying her choice of "64" in Q1a ("One Trial"), this is what she had to say:

> DS: Because, odds are, 60% are red, and you're probably not going to get exactly 60%, just because of the variability of the blind drawing… so 64 is close.

Thus, DS expects variation in her draws from the Large Jar. With the spinner in Q12, she said:

> DS: And so, I think as you're , like, close in the area of 25, which this little box and whiskers says, you know, close is between 22 and 29, which isn't on either end very far from 25.

Note how she is influenced by the boxplot, whose distribution for the middle 50% of the data ranged from 22 to 29.

In summary, DS does think in terms of what is likely or unlikely, and she also uses proportional reasoning as a justification. She is influenced by how data gets distributed, and expects variation in random situations. The next aspect focuses on displaying variation.

Displaying Variation

The questions that informed this aspect were chiefly Q8 (“35 Muffins”) and Q9 (“Two Bakeries”), and to a lesser degree Q5 (“Small & Large”) and Q12 (“Compare Comments”). All four of these questions included graphs in the presentation of data. First the dimension of evaluating or comparing graphs will be discussed, followed by the dimension of making conclusions about graphs.

Evaluating or Comparing Graphs: The first category in this dimension concerns attention given to the center of the distribution. In the case of DS, she did specifically attend to the mode when looking at the graph for Class B in Q5, and this was a part of her comments shared earlier about expecting less data the farther away the data is from the expected value (which was the modal bin of 61-70 reds for that graph). DS also focused on the mode when referring to the histogram in Q8 and to the median when looking at the boxplot for the same question. Attention to centers does usually tell much about how a subject is reasoning about variation, but it is an important feature of basic

graphicacy. More importantly, unless a person can reason about centers, it is unlikely to discover how that person reasons about data being spread in relation to that center. Therefore, it is useful to note not only that DS comments on the measures of central tendency, but that when applicable she can do so with different graph types as in Q8.

Going beyond centers, DS showed good attention to elements of the distribution when evaluating and comparing graphs. For example, in Q5 ("Small & Large"), she showed sensitivity to the number of data points at the extremes for Class B at the Large Jar. In Q8, she looks at range and interquartile range, in addition to the mode:

> DC: [Q8b] Are there any similarities or differences in the way that these two graphs show you the data?
> DS: Um, well, this one {Boxplot} you can see more… I think it {shows} real clearly that 50 % are really clustered between 112 and a half and 115 and a half, and so, you go "Oh, most of 'em…you know, the middle 50% , have a very small range of weight" and here {Histogram} you're kind of like, "Oh, like there's a big range of weight, but there's a lot in 113"

Her response shows that she was looking at the IQR in the box plot, and notice she uses a good adjective ("clustered") for describing relative placement of data. She contrasts the range for the middle 50% of the data shown in the boxplot with the overall range shown in the histogram, and also brings out the way that the modal muffin weight is 113 (a feature not revealed in the boxplot).

Further language connoting distributional reasoning comes through as she explains why, in Q9 ("Two Bakeries"), she thinks the East End Bakery is

not very consistent:

> DC: [Q9] What makes you think that?
> DS: Well, because they have this really big range for their high and their low weights {She is looking at the Boxplots}, and um, and also their… Here {the Dotplots} their plots are scattered, you know, further apart, and you're kind of – A little here, and a little there. And here {West Dotplot} you're kind of, grouped in a nice , you know, like most of them weigh right in the middle here, and then there's just a couple little ones that, you know, accidentally weigh less <Laughs>

Again, DS attends to multiple elements of the distribution while looking at both graphs. She looks at the range, and also where "most of them" (the data points) are. Moreover, she uses words like "scattered" and "grouped," terms that are descriptors of the distribution. She continues to expound not only about the extreme values, but also the upper quartile of one set of data:

> DC: What does it show you?
> DS: Well, it looks like your low weight here really is more of an outlier because it's like, almost the same length as your whole middle 50%, and that your higher end 25% is still real close to your middle weight {She's looking at the West Boxplot}.

Finally, even in Q12 ("Compare Comments"), DS gave responses showing how attended to the distribution of data in both the dotplot and the boxplot. She said that the dotplot "is kind of a bell curve" and that "the box and whiskers shows everything's kind of clustered right around that 22 and 29." Taken together, the responses of DS show a rather high level of reasoning when evaluating and comparing graphs, because of her synthesis of multiple characteristics of the distribution.

In terms of levels of detail and the perceived usefulness of the different

types of graphs, DS noted in Q8 that she could better see the variation in the data when looking at the histogram as opposed to the boxplot. I then asked:

> DC: How does it {the Histogram} do that for you?
> DS: Well, because it has each thing detailed out, so you can see how many are exactly which weight, where this {Boxplot} gives you the general range for you know, the percentage of the numbers

Her comment agrees with the fact that boxplots give a coarser view of the data by using quartiles, whereas the detail in a histogram is linked to the bin widths (which in Q8 were in relatively fine increments of 1 gram to capture the weights of the muffins). By Q9, DS expressed a preference for the boxplots because she was using the quartiles to draw a comparison between the two bakeries.

Making Conclusions: DS made a few conclusions regarding the graphs in Q5, Q8, Q9, and Q12. By conclusions I mean that it was apparent she had thought things through and had reasons for what she thought. For some questions, these conclusions sound tentative at first, such as in Q5 ("Small & Large"): "Well, I would think, if somebody made it up, it would probably be {Class} 'B'." A few sentences later, DS confirmed that "I think A looks more real than B." In Q8 ("35 Muffins"), the main conclusion I could gather was how DS thought that if she had bought a 36th muffin from the bakery that had produced the 35 muffins whose weights she had been analyzing, then she would expect her muffin to weigh about the same as the median weight of the other muffins. The way in which DS took into consideration the mode, as well

as the median and the IQR, showed me that DS's expectation was carefully considered, and that is why I have listed it as a conclusion. In Q9 ("Two Bakeries"), her assessment was: "I think that the East Side bakery is not very consistent with their weights of their muffins." Finally, in Q12 ("Compare Comments"), DS decided that she was comfortable with the data, and that the graph showed accurate results instead of revealing a flawed or biased spinner.

Interpreting Variation

The main dimension of the aspect that came through in the responses of DS had to do with influencing expectation and variation, and the questions that elicited these responses were Q1, Q10, and Q12.

Influencing Expectation and Variation: A number of comments made by DS show that she considers the questions with in the context of the number of trials. For example, in Q1 she says "I think with only six drawings, you'll probably stay close to the average," and this suggests that if there were more than six drawings, results may stray further from the average (she uses the average to mean the expected value of 6 reds in this situation). Conversely, when using the spinner, DS notes that "the more times you spin it, the more chance that you'll get 25," meaning that her expectation of getting the expected value also seems to rise as she does repeated trials. Her language was in terms of doing more spins, but a trial is actually 50 spins; however, I believe the idea that DS was trying to impart was that more trials affords more opportunity to obtain a result of 25 blacks out of 50 total spins per trial. She also mentions that "on average it'll be close to 25," again implying the

repetition of many trials on which to base an average. Later, in Q12, when asked if the results look about what she would have expected, DS first takes into consideration the previous question on the PostInterview, Q11, which had different lists of hypothesized results coming from six trials at the spinner:

> DC: [Q12] Is that about what you would have expected?
> DS: Well, this is , you know, different from the other page, but sure, that could happen, yeah…Oh, this is more spins, yeah, that looks good. I like this.

Her comment about Q12 being "more spins" had to do, I believe, with the way she was rationalizing about the results for six trials in the previous two questions (Q10 and Q11), and then moving the increased number of trials – twenty trials – in Q12 allowed her to be comfortable with the extreme values of 16 black and 34 black shown in the Q12 results. All of the above examples show some connection between number of trials and either expectation or variation, and as a final example, we have DS being very explicit:

> DS: [Q1] And the more times you pull, you'll have variations on each end, which might get wider, but you'll have more in the center, around the 60

This is, I think, a very reasonable interpretation of variation that DS makes.

Discussion

Since three of the questions used in the studying the cases' responses on the PostInterview had parallel questions asked in earlier instruments, and the each of the three questions relate to a separate context (Q1 relates to sampling, Q9 relates to data and graphs, and Q10 relates to probability), it will

be useful to look at the thinking of DS across the instruments on these related questions.

On Q1and Q10 in the PostInterview, DS does not choose the expected value as her guesses on “One Trial”, while on the similar questions in the PreSurvey and PostSurvey she did choose the expected value. However, analysis of her earlier responses to *why* do show a consistency in that she knows the results could be more or less than the value she has chosen. For example, in the PreSurvey she wrote that she “could get more or less” (PreSurvey Q1a), and in the PreInterview she noted that the prediction of 6 reds was “not for sure” (PreInterview Q1a). DS also shows consistent reasoning in both the sampling and probability contexts when asked to predict results and offer reasons. There is another interesting comparison between the two contexts on different instruments when reasoning about graphs, and to share this I will need to introduce responses from two other questions from the PostInterview.

In the sampling context, PostInterview Q3 (“Real: 30”) shows data supposedly resulting from 30 trials at the Large Jar, and compares well with PreInterview Q3 (“Fake: 30”) at the Small Jar. Whereas in the PreInterview Q3, results only occurred on 5, 6, and 7 reds (they were handfuls of 10 from the Small Jar), in the PostInterview Q3 the results went from 53 to 69 (they were handfuls of 100 from the Large Jar) and had a mode at the expected value of 60 red. In the PreInterview, the graph for Q3 was made up but DS declared it reasonable, suggesting that those people doing the 30 trials just

happened not to get much variation. In the PostInterview, the graph for Q3 was real, and DS also declared it reasonable, but her reason had more to do with the presence of perceived variation:

> DS Well, because our, you're most common number is 60, which is the average number of reds, and then, there's kind of a cluster around that number. And then there's just a few on the edges, you know, a little straggle here and a straggle there {She marks the min and max}

Here, DS saw the maximum and minimum values and decided that they were far enough away from the "cluster" to agree with her notion of not being too perfect, which is exactly the problem she had with PostInterview Q13 ("Real or Fake?"). Set in a probability context, PostInterview Q13 shows two graphs, each showing the supposed results from 30 trials at the spinner (50 spins per trial). Class A shows fake results which have a reasonably bell-shaped appearance, ranging from about 22 black to 28 black, and this class was quickly denounced as a fraud by DS. When asked why, she said:

> DS: Well, because it's toooo perfect. There's … 'cause… and, You know, that… I think, I just think somebody would have gotten, you know, under 20, or 20, or you know, 32… Something with a little… You know, even if it's one person, that would be out of the cluster. {Showing with her hands…wider spread?}

Her response to this final question on the PostInterview shows a shift from her reaction to Q3 on the PreInterview, which also had too little spread.

Finally, regarding Q9 ("Two Bakeries"), DS's reactions were very similar to that on the similar question on the PreInterview, which was also numbered Q9 ("MAX Wait-Times"). Namely, DS was quick to use language suggesting

relative tightness of the data, and also the implications for making judgments about consistency or reliability. While DS did not volunteer reasons for the variation in the weights of the muffins in “Two Bakeries”, she did have some ideas about why there would be differences in the “MAX: Wait-Time” data sets.

To summarize, DS has some reasonable ideas about variation. She draws from her prior knowledge and experience to accurately identify likely and unlikely expectations, and knows that results vary from trial to trial. While at some times she emphasized that more trials gave more opportunities to get extreme values, at other times she noted that more trials also gave more chances to get the expected value – Both are true observations that she makes, and it shows how her attention goes to both the ends of the distribution as well as to the average. Her ability to read graphs seemed to be a significant help to her as she evaluated and compared data sets in graphical form. That is, she could reason from histograms, dotplots, and knew what the boxplots were telling her about how the quartiles were spread. At the end of the PreInterview, given the way she had erroneously identified Q3 (“Fake: 30”) as showing genuine data, I had thought that more experiences actually doing experiments similar to the situation in the question would be of benefit. It is interesting that on the similar questions asking for real-versus-fake judgments on the PostInterview, which was administered after the in-class simulations, she was able to correctly identify the genuine graphs and give a reasonable explanation.

The Case of GP

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

Throughout the PreSurvey and PreInterview, many of GP's responses showed more of a naïve kind of thinking about probability and statistics. For example, when telling why repeated trials at the Small Jar wouldn't be the same, GP wrote on the PreSurvey that "you will probably grab differently and the candies are shifting to different places." In the PreInterview he repeated this theme, and actually offered a physical demonstration of how "your hand creates randomness" by grabbing differently each time. He suggested in the PreSurvey that more trials gives "more chance of getting a radical number," which was his language for extreme values or outliers. In fact, during the PreInterview, when commenting on different lists for possible outcomes on six trials from the Small Jar (PreInterview Q2 "Lists"), GP liked a list containing 1 red and 10 red. In choosing that list, he said "I like that. I did that because I kinda wanted to be a little radical {Laughs}." The latter quote tells more about

GP's character than anything else. Other responses showed GP to have some good intuition about variation, and I noticed that these responses particularly showed up in questions concerning graphs. For instance, in the PreSurvey question concerning 50 trials at the Small Jar (Q3 "Fifty Trials"), GP had written that the "top of the pyramid is 6 or the most probable and it just cascades down." He repeated this theme in the PreInterview, using his hands to show a tapering off from the mode, and it became clear that he had some kind of bell-shaped distribution in mind. He also was mindful of distribution in the PreInterview Q9 ("MAX Wait-Times"), when he was torn between the data sets having the same summary statistics yet different amounts of variation.

Expecting Variation

The responses which illuminated this aspect for GP came from the same questions as were used by DS. That is, all of the six questions except for Q9 ("Two Bakeries") helped inform this aspect.

What was Expected: GP expresses that results should be "around" the expected value, and he uses this term in relation to both the sampling context (such as in Q1, using the Large Jar) and in the probability context (such as in Q10, using the Spinner). However, as will be described in more detail in the *Interpreting* aspect, GP also mentions that more trials means results will be closer to the expected value. An extension of his thinking seems to be that the less trials done, the further your results could be from the expected value. Thus, on Q1a ("One Trial"), a guess of 70 reds is what GP sees as reasonable

for his first trial. On Q1c ("Pick Six"), he put "48, 50, 58, 62, 68, 72," saying:

> GP: I just thought that 60 is kind of where it's going to end up if you grab a lot, you want to pick around 60, kind of going a little bit more extreme, 'cause of the…Well, I mean, I just – Picked around 60, and 10 or whatever {Waving his hand on either side of the graph}

The above comment shows some of the difficulty in picking apart what exactly GP means. For instance, when he says "grab a lot," it's not clear if he means grabbing larger handfuls (of 100 in the Large Jar as compared to 10 from the Small Jar) or making lots of grabs, as in repeated trials. Also, when he says that he "picked around 60, and 10 or whatever," GP's hand-waving and previous answer of "70" on "One Trial" convince me that the expected value of 60 reds (plus or minus 10 reds on either side), is an adequate definition of "around 60" for GP in this situation. For "One Trial" at the spinner (Q10a), GP put 20 blacks out of 50 spins, and when asked to compare possible results on a second trial (Q10b), GP said "It could be higher, it could be lower. But, you know, I would guess maybe 27." For "Pick Six" at the spinner (Q10c), GP put "21, 23, 24, 25, 27, 28," and when asked why he simply said: "Around 25."

Two other responses in different questions also showed the kinds of results expected by GP. In Q8 ("35 Muffins"), when asked how much his own muffin might weigh, GP said "around 113 grams." The value of 113 grams corresponded to the mode shown in the box of summary statistics as well as on the histogram. Lastly, when looking at the graphs in Q12 ("Compare Comments"), GP showed support for Keith's comment of suspicion about

there not being any 25's (the expected value) in the data set. Graph does show results that might be construed as close to 25 (there is one data point at 24 and one at 26), and GP said

> GP: You know, you {Referring to Keith as if he were there} bring a good point, you would think that you would hit a 25, but you're spinning this 50 times, and uh – You know, you're going to have some , you're going to have different counts.

As will be discussed further, GP agreed the results were reasonable, which is in accord with his sense that results should be around or close to the average or expected value.

Notice in GP's latter comment how he expects results to be different from one another. Although he was not as explicit as some of the other subjects about not wanting to see repeated results, it seemed clear that his main expectation was to see differences. For example, in Q1b ("Repeated Trials"), GP commented that "very likely {you're} going to get a different group, a different combination." Also, when talking about how he liked a list of numbers for a "Pick Six" scenario, he said they were "just pretty random numbers." This was on PostInterview Q2, which although not part of the common subset for all cases, is useful to introduce at this point because of what GP has to reveal about his thinking on randomness. Q2 had lists of possible results on six trials from the Large Jar, analogous to the PreInterview Q2 for the Small Jar.

> DC: What is it about the list that makes your mind think "just random" numbers?

> GP: Um, there all different, there's no rhythm to 'em… Mmm-hmm. 'Cause the other ones {Lists} that I don't like have rhythms to them.

Thus, it can be seen that GP has a preference in expectation for results to look different, and it is therefore natural that all of GP's own "Pick Six" guesses had no repeated values.

Why (Reasons for Expectations): The first category of reasoning *why* that I observed in GP's responses had to do with likelihoods and possibilities. He sees the extremes as possible, and concurs that there is no guarantee of getting the expected value in sampling or probability contexts. Again, these types of arguments or observations are listed as reasons *why* when they are offered as justification for whatever expectations the subjects have posited. For example, in discussing Q1 (using the Large Jar), GP at first made it seem as though the extremes were likely by saying "you're more likely to get more of the, I don't know, extreme numbers." He then went on to clarify that he meant he thought it was likelier to get those extremes in the Large as opposed to the Small Jar, and this idea will be discussed further in the *Interpreting* aspect. He was clearer when talking about his choices for Q1c ("Pick Six"). I asked why he hadn't put any numbers in the 20's or any 5's and he said that my suggestions were "less likely to do {get}. So I just picked more likely options." Also, in Q5 ("Small & Large"), he commented on the graph for the Small Jar that "I think it's pretty hard to get these 2 and 9's." It became clear that GP had an idea that extreme values were unlikely, but he

also emphasized the possibility of such outcomes when talking about the data set in Q12 ("Compare Comments"):

> DC: Ok. How about Jeanette's argument? She's disturbed with the maximum of 34…
> GP: Mm-hmm. Well, as you can see from us doing this, it's possible to get a high number. You know, is it possible to get 36? Could it happen? Sure! Sure it could happen. It's very unlikely, but people get lucky in life. You know, sometimes you'll win the jackpot, sometimes you don't. {Laughs}

As a final observation in this category of reasoning *why*, GP notes that it's possible to not get the expected value, and even probable to not get it in multiple trials, when further discussing Q12:

> GP: Isn't it possible not to get to a 25, when you do the spinning? Out of these 20 students, {they} just happened to not get a 25! And, as we see, it's possible {of} that not happening.
> DC: Ok
> GP: You know, I bet – If we did this experiment again, is it probable that we might not get another 25? Sure

Moving beyond softer references to what might or might not happen, GP also was able to use proportional reasoning in justifying his thoughts about the expected value. In some cases, it was important to not take GP's proportional reasoning for granted, since in the PreInterview he had seemed to be heavily influenced by the numbers of candies in the jar. With Q10a ("One Trial"), he had first volunteered a guess of 45 blacks out of 50 spins (which seemed very extreme), but then he gave 20 blacks, after having said:

> GP: Oh wait, no! What am I talking about? I gotta divide it by two… Well, 25 would be half, and 20 is possible {Laughs} Right, right… The middle is 25… That's the odds, the higher odds of getting 25

His other form of reasoning *why* came through as he referred to previous experience as a way of justifying his thoughts. For instance, talking about how it was "pretty hard" to get the results shown for the Small Jar in Q5 ("Small & Large"), GP added "I mean, really hard. I mean, we did that in class – what was it? Did we do something like this in class?" Later in the PostInterview, when mentioning how he could convince others of what results in Q12 ("Compare Comments") should look like, GP said: "And that's when I would pull out the Phantom {Fathom} software and show 'em how this works." In both of GP's references to previous experience shared above, he was talking about specific activities or interventions done in the Math 212 classroom. We had indeed done sampling from Small and Large Jars, and we had also used the Fathom software on more than one occasion to replicate large numbers of trials of experiments done in class. GP was using what he recalled from those class experiences as reasons for his responses to some of the questions in the PostInterview, and this is an important point which will again be emphasized in the concluding chapter.

The last reason *why* offered by GP had to do with the presence of variation, and even though it only came up in one question of the subset (Q12), it is worth mentioning because it appeals to some sophisticated notions:

GP: Ah, 20 sets of 50 spins… This is only 20 kids, that are doing this, um… You're going to have a lot more variation in where the median and the mean are going to go.

DC: So with the 20 sets, though, you could imagine the mean the median moving around from 25, is that what you were saying?

GP: Sure, sure. It could be moving around the 25, the mean…
DC: Like from where to where, do you think?
GP: Well, just looking at the boxplot, you can see it, you know, from 22 to 29, you know. But, um, it would probably move around 20 – Yeah, around that area, it could be moving around

Contrast the above comments with, for example, what RL had said in the PreInterview about how individual results might vary but the average should be the expected value. GP explicitly expects variation not just in the data points, but in the resultant mean and median of the data set. I was wary that GP was using the terms of central tendency without quite making sure of what he was suggesting, and that was why I asked further questions. At the very least, it seems that GP is more open than some of the other subjects to expect averages to not always match what the ratio for one trial would predict.

Displaying Variation

Once again, the questions that informed this aspect were Q8 (“35 Muffins”) and Q9 (“Two Bakeries”), and also Q5 (“Small & Large”) and Q12 (“Compare Comments”).

Evaluating or Comparing Graphs: GP gave fewer comments that directly showed he was focusing on the average of a graph, but gave more indication that he was influenced by the range or distribution of the data. In Q8, when predicting “around 113 grams” for the weight of his muffin, GP mentioned that the median shown in the boxplot and listed with the other summary statistics had persuaded him. Also, in Q12, when he talked about

the “median and mean” going to different places, again he was attending to the central values on the graphs.

With regard to range and distribution, GP recalled the basic idea that boxplots were showing the data broken into roughly equal quartiles, and he appealed to the boxplot in Q8 when talking further about his muffin weight prediction:

> GP: Ok, um… So the majority would be in here, is that what it is? {Pointing in the IQR, around the median} Ummmm. About 50% would be this {Encircling the IQR}. So I’d just say the interquartile range, you know… From this {Q1} to there {Q3}

Thus, he expanded his prediction when evaluating the graph, and implied he’d be comfortable expecting a muffin that was somewhere in line with where the middle 50% of the rest of the data fell. In Q9 (“Two Bakeries”), compared both data sets using both graph types when he was explaining why the West End Bakery seemed more consistent:

> GP: Well, you can just look at the boxplot here… The middle 50% is a lot shorter than the middle 50% of the East End Bakery. You see down here, in the dotplot, that there’s quite a bit of difference where the dots are, little groupings where the dots are. The West End Bakery is more closer together.

In the above comment, GP combines several features of the distribution. He compares the IQRs of both boxplots, and then also compares the dotplots, using expressions like “little groupings” and “more closer together” to show how he sees the data distributed. He later contrasted the West End Bakery to the East End, noting that the latter was “more spread out.” As a final

example, GP again compared both graphs at once in Q5 ("Small & Large"), and showed an implied focus on the shapes of the distributions:

> GP: Well, I'm just seeing that they kind of have the same kind of ratio / height kind of thing, {His hands are tracing on both graphs at the same time}

Again, the language that GP uses is somewhat nonstandard, but in referring to the "ratio / height kind of thing" and using his hands to show me what we meant on the graphs, it seemed clear that GP was attending to the similar shapes shown in graphs for both the Small and Large Jars. All of the above examples portray different degrees of reasoning about some aspects of the distribution, showing that GP considered more than mere centers when using displays of variation. Another consideration in evaluating or comparing graphs is the amount of detail shown and subsequent usefulness of different graph types, and GP noted how the histogram offered more detail than the boxplot in Q8 ("35 Muffins"):

> GP: You see the different altitudes of numbers here {Histogram}. See both the ranges, right? You see an empty spot here and here {Histogram 110.5 & 116}, but you don't see empty spots in this {Boxplot}

On the surface, GP's comment may seem difficult to interpret, but his assessment, while fundamental, shows good appreciation for the differences in the two graph types. In particular, GP was noting in the histogram that he could see some "empty spots" where no muffins had attained that weight (such as 110.5 g and 116 g). However, since a boxplot by its very nature obscures details, he could not see "empty spots" because boxplots appear

with continuous quartile. GP went on to note that "here {Histogram} you see more variation, 'cause of the ups and downs of the graph" which again is a telling remark about how GP sees variation in the different graphs. An important note at this point is the shift from equating variation with range, as is typical for novice learners of probability and statistics, but GP specifically equates variation with "the ups and downs" that he sees. It was interesting to then go back and see what GP put on PreSurvey Q6, when the question asked which graph (School A or School B) showed more variation. Since the vertical variation seems more pronounced in School B, I would have thought GP might have selected that school as having more variation, yet in fact he answered correctly by giving the school with a wider range (School A). Lastly, although GP did seem to use boxplots quite readily, he expressed the opinion that others may prefer dotplots in Q9 ("Two Bakeries"):

DC: Yeah. You know, is one graph more helpful than the other in showing you the differences between the two bakeries?
GP: I think people would understand this one {DotPlots} more than this one {BoxPlot}. You'd kinda have to be taught to know about this one {BoxPlot}.

Again, GP's point is well expressed and accurate. For rough comparisons of distributions of data, a boxplot may suffice, but showing finer degrees of variation is better served by another graph type, such as a histogram or a dotplot.

Making Conclusions: As mentioned earlier, in Q9 GP did say that "you see that the West End Bakery has a lot more consistency in their… the way

they make their muffins," and this conclusion was influenced by the way GP had seen the data as being "more closer together." When he compared the graphs for the Small and Large Jars in Q5, GP concluded that Class A (with the Small Jar) had made up the data. He was suspicious of the way the two graphs looked so similar in shape:

> GP: Um, it {Small Jar} is very similar to this {Large Jar}, like maybe they saw this and copied it, kind of, and just….You know, it looks like it's almost, like, you know, they went directly and just did, like, you know, {Shows a sort of copying from Large to Small graphs} just made it a little shorter and stuff, and just…maybe a bit taller

In fact, the graph for the Small Jar shows real data, and so the idea that the graph for the Large Jar shows too much variation does not get mentioned by GP. However, based on some of his other comments about how the increased number of candies makes it likelier to achieve extreme outcomes, his reasoning seems consistent. Finally, with Q12 ("Compare Comments"), GP concluded that the graph looked about what he would have expected, largely because it seemed possible to him for the results to look as they appeared.

Interpreting Data

Five of the six questions from the common subset elicited responses from GP about this aspect (Q8 "35 Muffins" was the exception), and these responses touched on the causes and effects as well as the influencing of expectation and variation.

<u>Causes of Variation:</u> As a matter of emphasis, GP seemed to stress the physical causes of variation, as he did in Q1 for the Large Jar by saying

"you're grabbing a whole different group, and every group has so many different choice… I mean, so many different options." Notice how his language starts off with the physical actions of "grabbing," as though if only we could grab the same group each time, then we could eliminate all the variation in the situation. However, as his earlier responses to PreInterview Q1 show, we aren't likely to grab the same group each time because "while you reach in, the candies move, all over the place." It's worth recalling another of this subjects' earlier ideas about how "your hand creates randomness," which gives good insight as to what GP is really focusing on as a cause of variation. Also, his comment for the PostInterview Q1 about "grabbing a whole different group" (from the Large Jar) compares well to another of his earlier ideas in the PreInterview Q1 about how "if you stick your hands in there randomly, you could just pick up any number, from 1 to 10" (from the Small Jar) . Again, the issue is one of emphasis, and his comments almost seem to imply that if you don't stick your hands in the jar randomly, you won't get random results. This is why, for GP, I have noted the theme of the nature of the candy mixing as having something to do with causes of variation in the sampling situations.

It was no surprise that GP also found causes of variation in Q9 ("Two Bakeries") by citing human reasons for why the East End Bakery was perceived to be producing relatively unreliable muffin weights. He suggested causes such as "miscommunication in the staff, you know. High turnover of employees and mixup of, or… human error, reading the recipes." For the

more consistent West End Bakery, GP thought that maybe "they have a formula that they stick to, and they usually hit, you know, the right amount." On a question outside of the common subset but still related to muffin weights, PostInterview Q7 ("Reasons: Muffins") showed the results of twenty measurements that a class of 6th graders had made to determine the weight of a single muffin. The twenty measurements were not all the same, and GP suggested the following as possible causes of variation:

> GP: It could just be human error. They're reading the numbers wrong, uh… What else? You know, maybe somebody snuck a bite of the muffin, they took a little blueberry off… {Laughs} You know? Started picking at it…

Rather than seeming facetious, GP actually shows a good understanding of the context of the problem in this situation, and articulates reasonable causes of variation. Human error is indeed a plausible cause of variation in the muffin weight situations of PostInterview Q7, Q8, & Q9.

Effects of Variation: In the common subset of six questions, the effect that GP mentioned came in response to Q1c ("Pick Six"), when I had questioned him about his choices. He started off by explaining that "it could go anywhere," meaning that results could fall "anywhere." Although further probing showed that GP had definite opinions about where results would be more likely to fall, I have categorized opinions such as "anything can happen" or "it could be anything" as an effect of variation. This categorization seems appropriate because the subjects who do express the "could be anything" opinion tend to hold the view

that a wide range of outcomes is possible because of the variation inherent in the situation. For example, in the Large Jar that GP was commenting on above, the presence of variation in a set of results tell him that "results could go anywhere."

Another effect for GP that did not show up in the common subset of six questions but did show up in other questions was that variation makes it difficult to decide about some situations. For instance, in PreInterview Q2 ("Lists"), there were lists of possible "Pick Six" choices from the Small Jar, and GP did not choose a preferred list, explaining that "because there are just so many different choices, and so many different choices in the other one, that it's kinda hard to decipher for me." It seemed that GP meant not that the scenario was hard to decipher, but that making a choice of a favored list was difficult. Also, in PostInterview Q3 ("Real: 30"), which showed 30 trials at the Large Jar, GP did not want to label the graph as showing made-up data. However, his reason was not based on how the graph did or did not conform to his expectations, but rather on how it was difficult to decide:

GP: I can't see how you can say that this is – you know, you look at it and go "No, this is a fake" You know? Um…{Pauses} I would say that they're… It's tough, I dunno…
DC: Well, you tell me what's in your mind "It's tough"…
GP: Well, because it's, you know: Anything is possible in … You can't say "Oh, no you guys did this wrong, you cheaters!" {Laughs} You know? You can't!

The above exchange shows how an effect of variation for GP is that some situations are hard to judge. "Anything is possible" in the sampling scenario,

and consequently, “it’s tough” for GP to determine if the graph shows made-up data.

Influencing Expectation and Variation: In the sampling context, a dominant theme of for influencing expectation and variation concerned the numbers of candies in the jars. That is, what he expects for results and how those results might vary depend to a certain extent on the numbers of candies present in either the population or in the sample. For instance, in Q1a (“One Trial”), GP took note of how we were dealing with the Large Jar as opposed to the Small Jar of the PreSurvey and PreInterview by noting that extreme results were more likely to occur “since there’s more choices.” He also emphasizes the numbers of candies in commenting about the likelihood of getting different results, saying

> GP: You know, if you have thousands of candies, and you pick a hundred, you’re gonna have… Very likely {you’re} going to get a different group, a different combination.

It was in Q5 (“Small & Large”) that his attention to numbers really came through, and GP was very clear that the Large Jar should have a wider range than the Small Jar:

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

Aside from the fact that this interpretation is contrary to what statistical theory

would suggest, the point is that GP clearly sees the sizes of the jars as having an influence on the range he expects. It should be mentioned that the graphs were designed to look very similar in shape to one another, but the scales are in fact different because the samples sizes differ. That is, Class A (Small Jar) has results that range from 2 to 9, while Class B (Large Jar) has results ranging from a possible minimum of 21 to a possible maximum of 90. So, GP is implying the use of proportional reasoning in a graphical sense when he talks about wanting to see a smaller range in Class A than in Class B. That is, strictly speaking, the range for Class A is (9 – 2) = 7, while the range for Class B is a maximum of (90 – 21) = 69. What GP means is that he wants to see the graph of Class B look wider than the graph of Class A, since Class B is based on jar and handfuls containing more candies than Class A.

In addition to comments on numbers of candies in the sampling context having an influence on expectation and variation, GP also mentions the number of trials performed as being connected to the results obtained. With Q1a (“One Trial”), right after GP suggested “70” as his choice for one trial, he explained why by saying: “ ‘Cause the more you pick, the better chance of getting those extreme numbers.” Notice how the latter comments mentions the greater chance of getting the extremes, while in the probability context (Q10b: “Compare Trials”) he suggests that more trials gives a greater chance of getting the expected value: “It’s all in the spin, you know. The more you spin it, though, you’re gonna get closer to 25.” Also in Q12 (“Compare Comments”), it was shared earlier how GP had talked about how having “only

20 kids" doing trials meant that there would be "more variation in where the median and mean are going to go." He continued this thought by adding:

> GP: Um, you know, if we had a hundred students, your mean and median will probably get closer and closer – the more and more you do – you know, the closer and closer you would get to 25.

Now, although GP has said earlier in the probability context that more trials means more chances of actually getting 25, the idea in the latter quote seems more in line with the Law of Large Numbers. That is, the more times an experiment is performed, the closer the experimental probability tends towards the theoretical (expected) value. For the way that GP views this situation, the number of trials is seen as a way of controlling or influencing the variation and the average.

Discussion

Some GP's perceptions of variation seemed to shift over the quarter. For instance, in the PreSurvey for both context of sampling and probability, GP put down the expected value for both Q1a and Q7a ("One Trial" at the Small Jar and with flips of the fair coin). He also gave the expected value of 6 reds in the PreInterview Q1a ("One Trial" at the Small Jar). In those instances, he only offered proportional reasoning as justification:

> GP: [PreSurvey Q1a] 100 total, 60 Red, 40 Yellow. Take the zero away or divide by 10. 10 total, 6 red, 4 yellow.
> GP: [PreSurvey Q7a] 50-50 chance
> GP: [PreInterview Q1a] Well, there's a hundred candies in the jar, and I just took the zeroes away from the 60 and the 40, and just came up with 6 and 4.

However, for the analogous questions Q1a and Q10a (“One Trial”) in the PostInterview, GP does not give the expected value in either case. He is still influenced by the expected value, but speaks more in terms of what is possible or likely, and how results might be higher or lower. Also, contrast the differences in his choices for “Pick Six” on the PreInterview Q1c and the parallel question on the PostInterview Q1c:

GP: [PreInterview Q1c] {He picks 1, 3, 4, 5, 6, 10}.
DC: Why did you choose those numbers?
GP: You could just pick up any number, between 1 & 10, so…

GP: [PostInterview Q1c] {He picks 48, 50, 58, 62, 68, 72}.
DC: Why did you choose those numbers?
GP: You want to pick around 60, kind of going a little bit more extreme…

Although in the PreSurvey and PreInterview, GP’s “One Trial” responses were the expected value, on the his “Pick Six” choices on PreInterview Q1c were rather wide, and his reasoning was naïve (of the “it could be anything” type). On the PostInterview Q1c, while his “One Trial” choice of “70” seemed somewhat high, his “Pick Six” did not have the proportionally equivalent wideness as his PreInterview choices. More importantly, his reasoning on the PostInterview Q1c was an improvement over his earlier “it could be anything” type of thinking, in that he was picking numbers that were meant to be “around” the expected value of 60 reds.

While GP’s responses indicated some shift in perception on some of the sampling and probability questions, he showed more stability in his regard for variation in considering the graphs on PreInterview Q8 (“MAX Wait-

Times") and on PostInterview Q9 ("Two Bakeries"). Both questions were similar in that allowed for a graphical comparison of two different data sets with different amount of variation but similar or identical averages. In the PreInterview Q8, GP was very quick to voice his opinion:

> GP: Well, the Eastbound train seems to be more consistent. 'Cause their's seem to be more closer together {Pushing his hands together on the graph}, This {Westbound} seems to be less consistent, since it's more spread out… {Draws his hands apart}

The above response matched well with his thinking on the similar situation on the PostInterview Q9 with the "Two Bakeries." On both of the above two questions, GP was very deliberate in considering different aspects of the situation. For example, in "Two Bakeries", he looked at both the boxplot and the dotplot, and talked about how data was grouped in the different graphs. In "MAX Wait-Times", there were only dotplots to consider, but GP considered both the summary statistics and the shapes of the graphs:

> GP: Well, it's saying that they're the same in some ways, you know. They have both the same numbers in the top and bottom. But then, you know, they look different, down below. {Referring to the graphs}

It seemed that the graphical presentation of variation made a strong impression on GP, and that he was more facile in his explanations when he had graphs for which to refer.

In summary, GP shows that he has some reasonable perceptions of variation in terms of expecting, displaying, and interpreting variation. He feels that results in the sampling and probability contexts should be close to the

expected value, but not always the same result each time. He can compare graphs with respect to averages and spread, attending to the distribution of data to help him think about variation and its subsequent consequence on consistency and reliability. Also, he talks about the numbers trials in connection with a change in expectation in variation. Some notions still to be developed by GP have to do with how close to the expected value results might be in experiments involving different sample sizes or different amounts of trials. For example, his comparison of the graphs in PostInterview Q5 ("Small & Large") show that he has an opposite expectation of how sample population and sample size affect variation from what theory suggests.

It would be a mistake to take GP's relatively carefree style and nonstandard language as a lack of understanding about variation. He has some good ideas on which to build, and seemed to really connect with ideas when data was presented in graphical form. Also, some tasks, like the "Who Cheated?" dice-throwing question on the PreInterview (Q10), elicited some very strong (and reasonable) responses from GP. GP helps show me how asking questions in different formats and in different contexts can draw out conceptions that might otherwise stay masked. Finally, it should be noted that GP directly commented on some of the previous class experiences. He referred to the sampling activities done in class, and also talked about the computer simulations we had seen. It seemed to me that the computer simulation in particular had made an impression on him, and I suspect that

having more experiences of that nature might refine his conceptions of variation, especially in the sampling and probability contexts.

The Case of EM

EM was quite willing to share what she knew and didn't know, and she needed very little prompting to voice her thoughts during both interviews. She had taken Math 211 the prior quarter with Steve, and was clear about not having had any prior courses in probability or statistics. Looking ahead to the material we would be doing in class, she said she was "open to it" and "interested to learn more". Her response at the start of the quarter to what variation meant was that it had something to do with when "there is a pattern and something changes in the pattern," and she gave as an example the time in the morning when her dog awoke each day.

In the PreSurvey and PreInterview, it seemed that many of EM's responses were more typical of a student with a stronger math background (particularly in probability and statistics) than someone who had never before engaged in the learning of the topics. For example, on the PreSurvey Q1a ("One Trial"), EM wrote "5 or 6" in contrast to most of the students who only put "6" as a response. Also on the PreSurvey, her range for six trials (Q2a) went from 2 to 9, and then in her range for thirty trials (Q2b) expanded to go from 1 to 10. Finally, EM was able to identify the correct chances of winning on PreSurvey Q9 ("Two Spinners: 50-50 & 25-75"), while most of the other students were not. On the PreInterview, she talked about "things are random" when drawing from the Small Jar, and how that meant to her that she didn't

think one could get 6 red every time. She added that “I don't think it'll be for sure six, it's just I think you're more LIKELY to get six. That would be my average number” (more than the previous two cases, EM tended to emphasize certain words in her speech, and this emphasis was maintained in the transcribing process). She freely used words like “concentration” and “variation” and “spectrum” when talking about graphs that she was looking at in the PreInterview, also incorporating much gesture in reasoning about the graphs.

As a final way of offering insight into EM’s thinking prior to the PostInterview, there are two themes to which I’d like to call attention which appeared in the PreInterview and later continued in the PostInterview. The first is the idea of having a mathematical formula to figure things out. In the PreSurvey, EM had written as an explanation to a sampling question, “Sorry but I don’t know how to calculate these answers. I’m just going off instinct.” Also, when thinking about her “Pick Six” in the PreInterview (Q1c), she seemed very pensive and in want of a formula:

> EM: Ohhh. I don’t…I don’t know…{Big sigh}. Well, I think, like… I don’t know…I don’t have any set computation, but I think it’s somewhere around 6

At the time, my impression was very strong that EM believed a “set computation” might actually tell her what the results for six trials at the Small Jar would be. More statements of this type came through in both the PreInterview, and will be further discussed at the end of case study. The

second theme had to do with offering personal experience as a reason for her expectations. For example, in the dice-rolling question of the PreInterview, EM said "I just know that when I play games I usually don't get ones" and later, she offered a reason based on "instinct, thinking about when I've played games, and how often sixes came up, and how often fours…" She repeatedly referred to experience in the PostInterview, and this will also be discussed again later in the case study.

Expecting Variation

EM had quite a bit to say about what she expected, and the questions which helped her talk the most about this aspect were Q1, Q5, Q10, and Q12, with a small contribution being motivated by Q8. As with DS and GP, Q9 did not elicit any responses concerning this aspect.

What was Expected: EM expected results in the sampling context to be "close to" or "around" the expected value, and consistently expressed this view in the different parts of Q1 when pulling handfuls from the Large Jar. When justifying her "One Trial" response of "somewhere between 50 to 70 reds", EM suggested that

> EM: We can see that the numbers generally center around the same kind of percentage as there are reds to yellow , so 60% red, 40% yellow , so somewhere between 50 and 70%...

Notice how EM not only starts off by giving a range for her answer on "One Trial", but also invokes distributional reasoning in describing how results "center around" the expected value. Later in Q1c, for her "Pick Six" choices of

"54, 58, 60, 62, 65, 70", she noted that "they are close to the 60% probability, with the reds there. In general, you'll pull somewhere around 60." Her thinking in this regard carried over to the Small Jar in considering Q5 ("Small & Large"). She said, "from the small container, you know, with most of the pulls being around 5, 6, and 7, I would expect that." For EM, in the Small Jar "around" the expected value could easily mean 5, 6, or 7, and in the Large Jar, her "Pick Six" ranged from 54 to 70, again showing what it means for EM to have results "close to" or "around" the expected value.

She also had a consistent way of expressing these thoughts in the probability context, and in Q10a ("One Trial" with the spinner) she did not initially volunteer an expectation, but said instead

> EM: I think it will land there somewhere close to 50% of the time. I don't think it will always be 50% of the time, I think it will be probably between 40 and 60% of the time.

As I furthered questioned her as to what that prediction might translate to in terms of how many spins out of 50 the arrow might land on black, EM again gave a range, stating that "Out of 50 spins, somewhere between – what would that be? So, 25, between 20 and … between 20 and 30 spins." Then she affirmed that between 20 and 30 was "somewhere near 50%, right." Her above comments imply that results won't always be the same, so it comes as no surprise that EM's choices on Q10c ("Pick Six" at the spinner) were all different from one another. She had put down "18, 20, 23, 24, 28, 32," reasoning that those results reflected experiments done in class, where

"it was generally concentrated between, you know… 50% of the time, you know, somewhere near there."

EM's references to how results "center" or get "concentrated" reflect a characteristic of the underlying distribution. She also used a range - another distributional characteristic – not only in talking about what she expects for a single trial, but also for repeated trials. For instance, when asked in Q1b ("Repeated Trials") how the subsequent trials' results might compare to the first trial's, she said "I think you'll stay within that range." Additionally, having given a range answer on Q10a ("One Trial"), EM thought that a second trial's results "would be very similar," meaning that a similar range would be expected. An example of something unexpected in terms of range came from Q5 ("Small & Large"), when EM felt that the lower extremes for both graphs were too low. However, for the graph of Class B at the Large Jar, she agree that "I would expect between 50 and 80 to be where it is here," thereby giving the kind of range she was comfortable seeing in that situation.

It was not only in the contexts of sampling and probability that EM discussed ranges. She used some thinking about ranges and expectation when talking about the muffin weight she'd expect in Q8 ("35 Muffins"), saying "first I'm gonna expect it to weigh between 109 and 118.5." The numbers she cited almost correspond to the entire range of the data set, where the minimum weight was 109.5 grams and the maximum weight was 118.0 grams.

Why (Reasons for Expectations): When talking about her reasons for expectations, EM does uses general terms for what might or could happen, and what is likely to occur. Her thinking across all three contexts of sampling, data and graphs, and probability includes the idea that extreme values are possible but unlikely. For instance she talks about how "there will be times where you might pull more yellows – A lot more yellows – than red" on Q1b ("Repeat Trials" from the Large Jar), and this thinking translates into the possibility of having some low results in that sampling situation. Her remark seems to contrast with her thinking about the Small Jar in Q5 ("Small & Large"):

> EM: With 40 pulls it seems a little less likely that you would have some on the lower end – In my opinion, you might get a few more 9s, and maybe a 10.

She had a similar reservation about the lower extremes in Class B's graph on Q5, from the Large Jar: "And, actually, I would say the same thing for Class B…there's two or three that pulled between 21 and 30, and that seems a little low." With the spinner tasks, EM commented that "there's no guarantee WHAT number you're going to get." Consequently, she felt that the extreme values of 18 and 34 in Q12 ("Compare Comments") could happen. As a last example, EM also talked about what might or could happen in the "Two Bakeries" situation of Q9, as this exchange shows:

> EM: In the East End, you have – sometime you might get an 88 gram muffin, but you could get all the way up to 142. So if I was a big muffin eater, and I wanted to take my risk that I would get a nice weighty muffin, I would try the East bakery.

DC: Oh. Where might you get the lightest muffin?
EM: Same place. So it is a bit of a risk.

I found it interesting that EM used the language of risk here in this context, and it suggest how her expectations might be influenced by her reasons *why*. That is, she did not volunteer an expectation for what she might get for a muffin weight in this question (which called for a comparison of the two bakeries), but it can be inferred that she would expect something between 88 and 142 grams. Her language suggests that, while an extreme muffin weight might or could occur, such an event is seen by EM as unlikely.

EM's proportional reasoning skills showed in both sampling and probability contexts. In contrast to DS, whose language often put proportions in terms of odds, EM almost exclusively uses percentages. For example, instead of talking about how the odds of getting black on the spinner are 50-50, or having chances of one-half, EM noted that "black is 50% of the circle." Furthermore, the earlier quotes by EM show evidence of how she uses percentages to express her proportional reasoning.

The other category of reasoning *why* for EM has to do with prior experience. When she talked in Q1a ("One Trial") about why results would "generally center around" the expected value, she was basing her reason on "what I've seen from what we've done in class, when we pulled handfuls before." Similarly, she also spoke of results from the spinner tasks looking "concentrated" because that was how she perceived the experience "after having done it in class." Finally, although not in the common subset, Q2

("Lists" for the Large Jar) – which gave different lists for possible results from six trials – also elicited reasoning based on experience. Here is what EM said in response to List iii ("60, 60, 60, 60, 60, 60") and List vi ("30, 10, 90, 20, 60, 50"):

> EM: {List ii} Well, just from what we've done in class, I never pulled a number the same time, six times in a row.
> EM: {List vi} On choice (vi), I don't like, because it's too low, the 10 is too low. Um, from what we saw in class, you know, it took I think something like 500 tries before we got so low of a number

It seems clear from the latter response that EM's impressions of what is likely and why were influenced by the classroom experiences. Moreover, since the class only performed about 300 trials by hand, her reference is likely to point to the computer simulations we did, where larger numbers of trials took place.

Displaying Variation

From the common subset of questions on the PostInterview, the ones that helped inform how EM thought in regard to this aspect were Q8 ("35 Muffins") and Q9 ("Two Bakeries"), and also Q5 ("Small & Large") and Q12 ("Compare Comments").

Evaluating or Comparing Graphs: EM did not have very many responses that specifically focused on the average shown in a set of data. It seemed that although she did incorporate an average in her reasoning, centers did not dominate her thinking. Thus there were not repeated comments about what the averages were, or how they were different. An example of her attention to the mode occurred when she was reasoning about

how much her muffin might weigh in Q8 (“35 Muffins”). She was drawn to the mode of 113 grams that was shown in the histogram, saying that she was “looking over here at the histogram, and seeing that, um, 7 people saw it at 113 grams.”

Instead of the average, more of EM’s comments involved ranges or subranges, and looking at either the shape or distribution of data shown by the different types of graphs. For instance, in Q5 (“Small & Large”), she first looks to graph of Class A (Small Jar) and talks about expecting “around 5, 6, and 7,” which is small spread on either side of the mode of 6 reds. Then EM goes on to comment on both ends of the distribution, effectively saying that she thinks there should be less data on the lower end and more data on the higher on. She offers parallel reasoning on the graph for Class B (Large Jar), again looking near the mode and then to both ends of the distribution:

> EM: I would expect between 50 and 80 to be where it is here, and then 81 and 90, I would definitely think that , that seems alright to me, but there’s two or three that pulled between 21 and 30, and that seems a little low...

In the context of data & graphs, EM again stressed range in discussing Q9 (“Two Bakeries”), and in the Q8 (“35 Muffins”), she was quite facile at reasoning about comparing the distribution as shown in the boxplot versus the histogram. While lengthy, I am citing the following excerpt so that the attention of EM can be followed as it switches from the boxplot to the histogram and then back to the boxplot:

> EM: [Q8] Ummm, I'm gonna expect my muffin to weigh… I'm gonna go with the boxplot answer, of somewhere in the 50% - middle 50% range – I'm gonna expect it – and, plus, looking over here at the histogram, and seeing that , um, 7 people saw it at 113 grams {the mode}, and that it does seem within, like, 112 to 115.5 {she's circling a range of 112 to 115.5 on the histogram with her finger } it seems like that seems to be a concentration of data… So I'm gonna say – expect my muffin to weigh… First I'm gonna expect it to weigh between 109 and 118.5 {Still using the histogram} There's the whole range, and then I'm going to think that it's probably going to be in the interquartile range, of – um, like, 112.5 and 115.5 {now she's using the boxplot}

What has happened is that EM first discussed the "middle 50% range" – that is, the IQR on the boxplot – which happens to be from 112.5 g to 115.5 g. Then she looks to the histogram and notices two things: The first is that the mode shown on the histogram is 113 g, but the second is that the histogram confirms for EM how roughly half the data looks to be between 112 g and 115.5 g. She talks about a "concentration of data" in the histogram, while it seems that she is mentally matching this "concentration" with what she has learned about the IQR holding 50% of the data in the boxplot. After processing this boxplot-to-histogram comparison, she considers the entire range of the histogram (about 109 g to 118 g), and then narrows that back down as she reverts her focus to the IQR on the boxplot. I consider EM's process in this situation to be evidence of distributional reasoning as well as of good graphicacy. The narrative also convinces me of the important link between fluidity in moving between different representations of data and the ability to express reasoning about variation. As a final example of how EM expresses some facets of distributional reasoning about graphs, she talks

about the extreme ends of the distribution as well as a concentration of data in Q12 (“Compare Comments”):

> EM: It’s what I would have expected, I like the high of the 34, I like the low of the 16, the concentration is definitely higher than 25, but – I could see that happening
> DC: Where do you see the concentration?
> EM: 28, 29, 30 {She circles that grouping}. Although there is, 5 people spin, spun 22 or below {She counts the data points} oh, 7 people, but um… You know, 21 and 22 {Circle that smaller group}, and then 24 and 26, so, are right in there… I actually would, this is exactly the range that I would pick for 20 sets.

Again, EM mentions both extreme values (the 16 and the 34) as well as clusters of data that she observes within the range of values. Although there were no data points at the expected value of 25 blacks, she does look at the data for 24 and 26. More importantly, she looks for the “concentration” which she sees above the expected value at around 28, 29, and 30. She also sees the smaller cluster below the mean, and although she declares herself satisfied with the range, it is clear that she is comfortable with the entire distribution of data.

Regarding the level of detail and subsequent usefulness of the different graph types as far as showing variation, EM again used sound reasoning in comparing the boxplot to the histogram in Q8 (“35 Muffins”). She was very specific about the information each graph did or did not provide:

> EM: Well, this {histogram} gives me an idea of all of the data, whereas this {boxplot} gives me only the data, the 50 % - the inner 50%, and the {Pointing to the whiskers} you know, and then of course 25% on each side

Her comment shows how she is aware that data is pooled into quartiles on the boxplot, and then she continues to be more specific:

> EM: But I don't really kind of know exactly where the more data is, like: 112 isn't on the inner 50%, but I can see here {histogram} that, you know, two people got that, and so, that's the kind of information I can't get from the boxplot.

What she is referring to in the above comment is how the lower whisker, which ranges from the minimum of 109.5 g up to the first quartile of 112.5 g, passes through 112 g but does not indicate whether any data points really attained that value (since such detail is not a part of constructing boxplots). Even though she mentions "inner 50%" – usually a reference to the IQR – it seems more that EM's point is how the histogram does show in this situation how two muffins weighed 112 g while the boxplot does not. It comes as no surprise when EM picked the histogram as the graph that gave her more information about the variation in the data, and when asked directly, she said:

> EM: Yeah, I think the histogram does. Because it actually graphs out each, each time that a certain weight came up, and so I can see more variation there,

She then gave another example of how the histogram showed frequencies for more specific muffin weights than the boxplot. Her reasoning with two graph types followed similarly in Q9 ("Two Bakeries"), which involved boxplots and dotplots. In comparing the two bakeries, she seemed to equate having more consistent weights with having a shorter range:

> EM: I can see from both the boxplot and the dotplot that they {West End} are more consistent in their weights. Their weights are concentrated between 93 and 120, whereas in the East End, you

> have – sometime you might get an 88 gram muffin, but you could get all the way up to 142.

For this question, although she again see more specificity in the dotplot as opposed to the boxplot, because she is primarily relying on the range she is able to validate the usefulness of the way the boxplot displays the range in this situation:

> EM: Again, I like the dotplot just 'cause I can see exactly where each muffin's weight fell, although just glancing at the boxplot, I can see that the West bakery is more consistent because the span is smaller… or the range. I'm sorry, the range is smaller… That is more consistent, then.

In the above comments, EM clearly shows how she thinks that the West End Bakery produces muffins of a more consistent weight, meaning that the weights are closer together from the minimum to the maximum. She then uses the notion of consistency to make a conclusion based on the graphs, and this conclusion is shared next.

Making Conclusions: It is interesting how EM phrases her conclusion in the above "Two Bakeries" scenario. She says

> EM: If I wanted to go to a bakery where I had a good sense of what I was going to get, {where} they were more consistent in the weight of the muffin, I would go to the West bakery…

The meaning of the West End Bakery's "consistency" is that EM then has "a good sense" of what to expect. In contrast, she referred to the East End Bakery as having a larger range, meaning that she could get lighter of heavier muffins than at the West End, but she would be taking a "risk." Seen in the overall picture of expectation and variation, EM's conclusions in this regard

are quite credible, and more discussion follows when talking about the effects of variation in the next aspect of interpreting variation.

EM's other conclusions about graphs came from the contexts of sampling and probability. In Q5 ("Small & Large"), she reasoned that both graphs "look okay to me," and then proceeded to comment on how they also both showed more lower extremes than she would have expected. Her conclusion that she "wouldn't be able to tell which one made it up and which one didn't" does reflect indecision as much as it does a decision that she can't say for sure. I categorize this response as a conclusion because it seems to reflect the outcome of EM's reasoning about the graphs. Finally, as shared earlier, in Q12 ("Compare Comments"), EM felt about the data that "I would expect it to be somewhat like this," and in fact the range was exactly what she said she would have picked. Therefore her conclusion was that the graph showed data from an unbiased spinner.

Interpreting Variation

Through responses to four questions from the common subset of six for the cases, EM had one response concerning causes, a couple of comments about effects, and more than a few about influencing expectation and variation.

Causes of Variation: In a similar fashion as GP, only with less repetition or emphasis, EM also mentioned the physical causes of variation in the sampling context. For example, in Q1 at the Large Jar, she talks about

why you could get “a lot more yellows than red.” She gives a reason as follows:

> EM: Just because of there being the opportunity of, in the jar, you know. You never know what happens, a big bunch of yellows might be there and that’s where you reach, so. You’re shaking it all around, but…{Trails off}
> DC: Oh, ok. Alright. Big bunch of yellows maybe in the jar…
> EM: Mm-hmm. Ah, in one place, and that’s where a hand goes, and so maybe you pulled some more yellows...

EM envisions a pocket of yellows grouping together in such a way that as the hand grabs from that area, an unexpectedly high ratio of yellow to red candies results. Thus, she seems to see the nature of the candy mixing as a direct cause of the resultant variation, with an improper mixing technique yielding unlikely results. EM was also consistent in identifying the candy mixing as a source of variation, as she had said very similar things in the PreInterview when discussing the parallel question (Q1) for the Small Jar. For instance, when reasoning about her “Pick Six” on the PreInterview Q1c, she noted that

> EM: Occasionally, maybe some yellows got pushed over to the side, so you'll pull more yellow. You can't tell where they're going to fall in the jar. And so, I think that in some places, you're not gonna always have a red/yellow red/yellow, or, you know. So, in some places, there'll be yellows that have collected together.

I categorize the above responses as having to do with physical causes of variation because they point to the way the different candies literally fall into place as a result of mixing.

Effects of Variation: One effect of variation mentioned by EM was that "you never know what happens," as she was quoted earlier in response to PostInterview Q1 at the Large Jar. She used similar language in the PreInterview with the dice tasks, pointing out how the "luck of the dice" phrase "came from, you know, you just never know what you're going to get." However, this effect was not mentioned very much by EM, and instead she usually had some sense of what would happen, as she expressed in the "Two Bakeries" scenario. Recall that she had said she would have a better sense of what to expect at the West End Bakery, whereas a visit to the East End Bakery entailed more "risk." The effect of variation for EM in this situation is that she sees greater chances of accurately predicting muffin weights within a smaller interval for the West End Bakery, which correspondingly has reduced variation when compared to the East End Bakery. The latter bakery, having more variation, also is the place associated with more "risk" – Namely, the risk or chances of making a prediction that does not come true. I categorize EM's associations of increased consistency with reduced variation as an effect because it shows an outcome of the degree of variation present in a situation. Namely, she sees the likelihood of an accurate prediction being affected by the amount of variation.

Influencing Expectation and Variation: EM did mention a link between the numbers of candies in the jars and the likelihood of getting certain results in the sampling context. For instance, she explained her initial guess of

"between 50 to 70" in Q1a ("One Trial") by saying: "Because, there are 600 reds , 400 yellows, so there's more of a chance of pulling reds because there's more reds in the jar." This thought neatly matched with what she had said about results being "close to 6" for a single trial from the Small Jar in PreInterview Q1a: "Since there's more reds than yellows, you're likely to pull, you know, at least half will be red."

More of EM's comments which were directed at influencing expectation and variation had to do with the number of trials performed in sampling and probability contexts. Although she repeatedly referred to the number of trials in many responses, often it was hard to tell what difference she thought this made on the results. That is, she often called attention to the number of trials, but did so in a vague sort of way, as these two examples show:

> EM: [Q1c, "Pick Six" from Large Jar] You're only pulling six times, so I think – In general, you'll pull somewhere around 60
> EM: [Q12, "Compare Comments" with the Spinner] I could see it {no expected value among the results} happening, if you had 20 sets…

Also in Q12, EM reflected upon the Jeanette's concern about the maximum result of 34 black, saying "I could see that maximum happening, it only happens once , out of the 20 sets." The difficulty in interpreting the above remarks is that they do not suggest a definitive direction for how results might look with more or less trials. The comment for Q1c about "only pulling six times" seems to suggest that with more trials, perhaps results won't necessarily be "around 60". When the above responses are looked at in light

of other comments EM did make at various points in the PostInterview, the picture becomes a bit clearer. To give one example, consider how explicit EM was in Q1a ("One Trial"), when she stated that "the more you pull, the more opportunity you have for outliers."

Other examples come from questions on the PostInterview which are not part of the common subset, but are useful to help show how EM thinks with regard to the influence that the number of trials has on the results from sampling and probability situations. Q2 ("Lists") showed various possibilities for choices of "Pick Six" from the Large Jar, and EM shared these comments:

EM: {List i, "72, 91, 74, 63, 81, 78"} Right. (i) has some higher numbers, which seem – Out of six tries, less likely
EM: {List iv, "53, 41, 34, 60, 46, 52"} Choice (iv), I mean…I don't think it's really likely that on six tries, you would pull a 34
EM: {List vi, 30, 10, 90, 20, 60, 50"} On choice (vi), I don't like, because it's too low, the 10 is too low. Um, from what we saw in class, you know, it took I think something like 500 tries before we got so low of a number

There is a theme running through the above responses that suggests extremes values are less likely with a fewer number of trials, and more likely with a greater number of trials. EM was consistent on this theme as she looked at Q3 (Real: 30) and Q4 (Real: 300). In the following exchange, note how she first comments on Q3, and then she compares her reasoning to what she had said earlier for only six trials,

DC: [Q3, "Real: 30"] What do you think is most likely?
EM: I think the results are likely, except that I would think that maybe you would have a few over 70, or maybe one lower than 50 – In thirty pulls, so, umm…
DC: Ok. Now, on the earlier page, I'm pretty sure you were saying,

between 50 and 70 was your number one feel…

EM: I was, but that was only out of six pulls. And six, that was – I like that idea, but with thirty pulls, I think you're going to have more – um, chance for the numbers to be a little… More spread out.

She later goes on to talk about 300 trials in Q4, saying

EM: Um, I mean, this is… could be very likely, but again, with 300 pulls, I think I would see – I would EXPECT to see, maybe one 80, and maybe one all the way down to 40, I just would see more of a… {She draws attention to the extremes}

Finally, EM's thinking in this regard was consistent even in the spinner tasks of Q9 - Q13. She seemed to hold firm to the idea that an expanding range of results would necessarily follow from an increase in the number of trials.

Discussion

There are many indicators that EM has a reasonable sense of expectation when it comes to situations involving variation. While in the PreInterview Q1a ("One Trial" at the Small Jar), she said "my guess is 6," towards the end of the quarter on the parallel PostInterview Q1a ("One Trial" at the Large Jar) she gave an interval estimate which was appropriate to the sampling context. In the probability context, she again made plausible interval estimates for her own predictions, and also analyzed the data shown in Q12 ("Compare Comments") in a way that showed her appreciation for what actual results would look like. Her responses in the probability context on the PostInterview (using the spinners) were similar to those given in the same context on the PreInterview (rolling the die). For instance, while all the six subjects put a range of choices on the PostInterview Q10c ("Pick Six" at the spinner), only two students initially put a range on the PreInterview Q9

("Sixty Tosses" of the die) – EM and GP. She also showed sensible expectation in the context of data and graphs, when she was considering how much her own muffin might weigh in Q8 ("35 Muffins") or what weight of muffin she might get at the East or West End Bakery in Q9 ("Two Bakeries"). One way in which EM's sense of expectation could use some further refining is in regard to just how likely extreme values are to occur, and what the effect of sample and population size is on this likelihood.

Given how easily EM talked about her expectations in terms of ranges, it seems natural how her main way of talking about graphs also had to do with distribution of data, especially the range. She had a very clear way of talking about which bakery was more consistent in the PostInterview Q9, and she had a similar style when discussing the parallel question the PreInterview Q8 ("MAX Wait-Times"). For the PreInterview Q8, EM had said

> EM: Um, from looking at this, I would say that the Westbound trains are less consistent in their wait-times. There's more variance. So, you can be waiting 7 minutes, and you can be waiting 14 minutes. And then, the Eastbound trains are pretty consistent, anywhere from 8 and a half to 11 and a half minutes.

Notice the similarity in language between her response to the above PreInterview Q8 and the way she spoke about the bakeries in PostInterview Q9. In particular, a smaller range gets equated with a more consistent performance, and a bigger range implies a less consistent performance. In the PreInterview, the smaller range belonged to Eastbound trains, and in the PostInterview the smaller range went with the West End Bakery. It is

especially interesting that, in the PostInterview Q9, none of EM's explanations are based on the average, even though the summary statistics are provided. This is important because on the PostInterview, the two bakeries' averages were not identical whereas on the PreInterview, the two train lines' averages were all identical, having the same mean, median, and mode. In looking at the trains' summary statistics, EM had said

> EM: I know that the average says that {there is no difference in the averages}, but you also have two ends of the large spectrum on the Westbound train, and a shorter spectrum on the Eastbound train. So, even though the average wait time may not differ, the amount of wait time could be a lot less, or it could be a lot more on the Westbound.

What EM has effectively concluded is that she is not relying on the averages to inform her about variation – Instead she is relying on the distribution of the data as presented in the graphs. In the PostInterview, EM does not even refer to the averages from the bakeries. The East has a higher mode than the West (142 g compared to 114 g), and a slightly higher mean (110.75 g compared to 110.25 g). The West has a higher median than the East (112.5 g compared to 101.5 g). A case could be made that, on average, one might expect a heavier muffin from the East Side Bakery. Yet EM's comments have more to do with where she will find the more consistent muffin (the West) and avoid taking a "risk" (the East). She offered the same preference for trains in the PreInterview, saying

> EM: Well, I would rather be waiting for the Eastbound train, because I know "Hey, I'm either on that train somewhere between 8-and-

a-half and 11-and-a-half minutes", and I like that consistency.

Thus, EM seems to reason more from the basis of the range of data, drawing conclusions in the form of where results are more or less consistent.

There was a component of EM's interpretation of variation that was not revealed in the PreInterview but did come through in the PostInterview, and that component has to do with the effect of the numbers of trials on expected results. Although she made several comments in the PostInterview about this effect, there was virtually no mention of this in the PreInterview. One reason may be the impact of doing activities in class, since EM also shared more comments about what she had experienced in the PostInterview than in the PreInterview. Recall that she had given some references in the dice tasks of the PreInterview to outcomes she had recalled from having thrown the dice in past games. In the PostInterview, EM continued to refer to past experience, and in particular to results she had seen or activities we had done in class. It seems likely that the link between numbers of trials and expected results became forged throughout the Math 212 quarter.

As a final note of comparison, two examples were cited at the beginning of the case study showing how EM had suggested that perhaps the correct computations could definitively answer questions about expectation and variation. She had more comments of this nature in the PreSurvey and in the PreInterview:

EM: [PreSurvey Q1c, "Pick Six"] I just went from 3 up to 8. No formula, just guessing.
EM: [PreInterview Q9, "Sixty Tosses"] I'm sure there's a formula for

this. But, I don't know.

EM: [PreInterview Q10, "Who Cheated?"] I just think that will happen, I don't really have any scientific evidence for that. Or mathematical evidence.

The theme that comes through is that with enough math, one could know for instance what would be the correct expectation for sixty tosses of the fair die, or the correct choices for a "Pick Six" sampling scenario. What is interesting is that none of this form of reasoning came through in the PostInterview, and again this absence may have something to do with the impact of the class activities. That is, regardless of the theoretical expectations, for which some formula were derived or provided in class, actual results still vary. There is a nice connection back to something that EM had first put on the PreSurvey in about what was the meaning of "random" to her. She wrote that it meant "no rhyme or reason – There is no formula." Randomness and variation together make up the Janus of stochastics: Randomness looks to the domain of probability and variation looks to the domain of statistics, but they are still two faces of the same coin. For EM at the start of the Math 212 course, she sees the variation in the outcomes of random events and wants a formula. After multiple experiences with probability and statistics in class, she no longer mentions wanting a "set computation," suggestive perhaps of a more accommodating or accepting attitude towards variation.

The Case of JM

Although JM was very adept at sharing his serious thoughts about variation, he also flavored his speech with comments of levity, such as when

he mentioned bringing his “triple-beam balance” to the two bakeries in the muffins problem, or if the person doing trials at the spinner “wasn’t drinking the night before.” He seemed quite at ease during both interviews, was quick at expressing thoughts when he was certain, but pensive when he wanted to mull over a situation for which he was uncertain.

JM had taken Math 211 the prior quarter with Steve, and when asked on the PreSurvey if he had taken any prior courses in probability and statistics, JM wrote “no, not really. A little sociology,” which I assumed might have included a small amount of statistics. He described his own attitude going into the course in a positive way, saying it “sounds great, looking forward to it.” Each of JM’s interviews lasted longer than any of the other cases. A taste of how JM tended to be more expansive in his responses comes early in the PreSurvey, when he gave a more protracted definition of variation as “something that fluctuates and is somewhat unpredictable. There is variety or differences.” He then went on to give four separate examples of things that vary: “The weather, people’s attitudes, the shapes of rocks, snowflakes.”

There are three themes in JM’s thinking that I want to highlight before examining his thinking on the common subset of PostInterview questions. The three themes came through strongly in the PreSurvey and PreInterview, and had to do with proportional reasoning, getting different results, and physical causes of variation.

First, although all cases reasoned proportionally to some degree, JM was explicitly and repeatedly clear about his reliance on theoretical probabilities. The mathematical ratio was often at the forefront of his thinking in almost all situations involving sampling or probability. For example. JM put all tens in the PreInterview Q9 ("Sixty Tosses" of the die), and his explanation was tied to proportional reasoning. First he said "it's one out of six," and when pressed further as to why he thought his choice of all tens was reasonable, he added

> JM: Ok, the odds. I'm just thinking, straight odds. When I throw a six-sided dice, the chance of it coming up any one number are one in six. For each side. Ok. That much I know. So, but when I throw it more than once, I still know that the second time I throw it, it's still one in six.

JM's emphasis on how the ratio doesn't change with repeated trials shows in his above comment.

A second theme that came through strongly with JM was how differing results are possible, which I found an interesting counterbalance to his embrace of proportional reasoning. For instance, in the PreSurvey Q1a ("One Trial" drawn from the Small Jar), he had written that he would get "6, maybe," reasoning that

> JM: 60 out of 100 are red, so 6 out of 10 are red. The chance of pulling out more red in a handful is greater, but if it actually happens is left up to chance – Random.

He then wrote on the PreSurvey that six trials from the Small Jar could range from 0 to 10 reds, as could thirty trials – Any result is possible for JM, and this made it difficult for him to judge graphs showing false data.

The third theme for JM had to do with physical causes of variation. Although in a PreSurvey spinner question JM wrote that results depended on "how the Fair-Spin Gods look at these things," it was clear that he did in fact take the physics of the situation into account. For instance, he also wrote on the PreSurvey in regard to spinners that results "depends on the force used to spin, the resistance of the spinner, the direction of the spin…and other factors." More on the themes of proportional reasoning, likelihoods and possibilities, and physical causes of variation as interpreted by JM will be discussed later in the case study.

Expecting Variation

What was Expected: JM repeated the theme of being close to an average or expected value even more than the previous cases. His language for what he might expect in Q8 ("35 Muffins") was that his muffin might be "somewhere around" the median of 113.50 g, and this phrasing was very similar to what he'd used in the sampling context at the start of the PostInterview. For example, in Q1a ("One Trial" at the Large Jar), JM at first simply said "I would figure, I would think somewhere around the 60% mark. Approximately." He later gave a range answer which will be further discussed with his other references to ranges. He also reasoned about "Repeated Trials" in Q1b that "the distribution will be, you know, close to the mix of 60%," and again explained his preference on Q1c "Pick Six" of "52, 54, 55, 60, 64, 68" by saying that "they're close to the 60% mark." Clearly JM expects results close to the expected value in the sampling context, and this sense of expectation

carried over into the probability context as well. With the spinner tasks, on Q10a ("One Trial"), again JM's first response was "I'm going to say, he's gonna get close to 25 black and 25 white." He also again eventually gave more of range answer for one trial, and when asked to compare results for a second trial (Q10b), he said "I think {it'd be} fairly close in the sense that it's gonna be around the 25 {blacks}."

In Q12 ("Compare Comments"), JM was inclined to agree with Keith's argument about needing to see some "25s" in the results. JM said he'd expect "a few 25s" and that he would "expect 25 to be 'The Number'," which I took as meaning that the expected value should be prominent in the results. A natural extension of this view let JM also agree with Karen's suspicion about the mode and median not being at 25, and he said: "I feel the same way that she does, that it {mode and/or median} should be closer to 25." JM later was more explicit in indicating that he thought 25 would actually be both mode and median in this situation. What makes these comments interesting that instead of just saying that individual results or repeated results might be close to an expected value, JM suggests that averages of results should also either be or be close to that value. This idea ties in with the aspect of interpreting variation, since it relates numbers of trials to expectation and variation.

JM also expressed how results would be different from one another, and at times this came through implicitly, as in Q10b ("Compare Trials") when JM said about subsequent results that "you know, it's gonna be close to the first ones," implying to me that results would not be identical each time. He

was more explicit in Q1b ("Repeated Trials") in saying that "you're not going to get the same thing", and that "you can go lower and higher" in subsequent results. His own choices on "Pick Six" for the Large Jar (Q1c, "52, 54, 55, 60, 64, 68") and the spinner (Q10c, "21, 23, 25, 26, 27, 29") were in fact all different. A couple of questions outside the common subset also highlighted how JM thought results would be different and not repeated. In both Q2 and Q11, lists of "Pick Six" choices were given for the Large Jar and spinner, respectively. In Q2, for List iii ("60, 60, 60, 60, 60, 60"), JM noted that "it's just highly unlikely to get the actual number six times in a row", and he emphasized that because of the "randomness of the whole thing, you're going to get a different number almost everytime. ALMOST everytime." With the similar List iii on Q11 ("25, 25, 25, 25, 25, 25"), he said that it would be "highly unlikely to have the exact, you know, the experimental match the theoretical like that." It is worth remembering that this is the same subject who not only put all tens on the "Sixty Tosses" task of the PreInterview (Q9), but continued to support that choice as possible and likely.

Not only does JM expect results to be close to the expected value, and typically different results on different trials, but he does give examples of the kinds of ranges he expects to see. In Q1a ("One Trial" at the Large Jar), he eventually suggests that he might expect "45 to 75, somewhere around there," and in Q1b ("Repeated Trials") he says that

> JM: You're going to get somewhere close to that, within that, broad mix we had, between 45 and 75. I think it's a good chance of it

being within that range, whether it makes it or not, who knows? JM added a "plus or minus" flavor to his range expectations in Q10, such as when he suggested that "One Trial" at the spinner would result in "approximately 50%, but it will be , you know, plus or minus, maybe, 20% of that number – Somewhere in there." A literal translation of what JM had said would mean 25 blacks, plus or minus 5, thus ranging from 20 to 30 blacks, which is reasonable. He downsized his range for Q10b ("Compare Trials"), saying subsequent results would be at "25 blacks, plus-or-minus that 10% or so," and also in Q10c "Pick Six," where he explained how he liked his choices because "they're close to uh, that 50 percentile, that we're looking for , plus or minus – I'm thinking – 10% or so." JM's "plus or minus" theme also ran through his responses on expectation for Q8 ("35 Muffins"), as he first gave a median expectation of 113.5 g, and then added "plus or minus a half a gram, how does that sound?"

Why (Reasons for Expectation): JM frequently used the language of possibilities and likelihoods when talking about outcomes in sampling and probability contexts. For instance, despite at times about how he thoughts results would be different on repeated trials, he also said it was possible in Q1b ("Repeated Trials"). He also stated for "One Trial" in Q1a that "it's possible to get zero red, and it's possible to get 100 red." The notion that all kinds of results are possible made it hard for JM to decide about the two graphs in Q5 ("Small & Large"), as he observed that "they look pretty close, which is possible." Just as it was hard to make a decision about which graph

was likelier to have shown made-up data in Q5, it was hard to find much to disagree about in the data for Q12 ("Compare Comments"). After all, the low value of 16 blacks and the high value of 34 blacks were both possible according to JM. There was an idea that there should be some values at the expected value of 25 blacks, but again there was the qualification about being possible: "You know, it's possible, but I think we should have at least one 25." More than the other cases profiled so far, JM really seem to qualify any judgment with the notion that the results he was judging were possible. He also mentioned likelihoods, and some net effect of his comments is that some events (such as extreme results) are unlikely but possible to occur, while others (such as getting the expected value) are likely to occur but possible to not occur. Some of his reasoning on Q2 ("Lists" for the Large Jar) showed a combination of possibilities and likelihoods:

> JM: They're all possible, but likely I would say
> JM Choice (vi) {"30, 10, 90, 20, 60, 50"}, it's got, like, a lot of outliers. I mean, you've got 90, and 10… Which, when we go to extremes like that, um, they're highly unlikely and to have those – It's possible

He showed consistency in reasoning on the similar Q11 ("Lists" for the spinner):

> JM: {List iii, "25, 25, 25, 25, 25, 25"} Highly unlikely. {Long pause} But certainly possible.
> JM: {List vi, "30, 10, 45, 20, 25, 35"} I thought it was kind of unlikely that out of {six trials}, to have a 10 and a 45 like that, they just seemed too far out. Very unlikely in six spins {trials}, but – Possible.

The trend in JM's thinking is clear that extreme values in particular are unlikely

but possible, and this reasoning helps explain why he holds some of his ideas about expectation.

JM also relies on proportional reasoning, and tends to express his thinking mainly in percentages. For instance, he knows for the Large Jar in Q1 that "the mix is 60% red" and he knows for the spinner in Q10 that "the probability {of black} is 50%, because we only have two {sides}, we know it's going to be one or the other." In explaining how proportional reasoning led him the theoretical probability in the spinner tasks, he said

> JM: Well, we know that we have a fair spinner, and we know that it's cut in half. And 50% of the time , theoretically, it could land on one or the other. That means it's half and half. That's the theoretical. It's gonna happen, I mean, there's a 50% probability or chance of it happening.

Finally, in addition to his proportional reasoning and talk about possibilities and likelihoods, JM also gave a few references to randomness or variation in his explanations. In Q1a ("One Trial" at the Large Jar), JM stated that "we don't know because it's all random," and in Q10a ("One Trial at the spinner) he said that the person doing the spinning was "just going to have some variation" in the results. As mentioned earlier, there is overlap between *what* is expected and *why* in comments such as the latter one about variation. In the contexts of JM's remarks, he ties his predictions to the reason that there is variation in the situation. The link between variation as a reason and variation as an expectation also can be seen in the comments about randomness. For instance, in Q2 ("Lists" for the Large Jar), JM notes that "it's so random, and

the fact that every time you go in there, it's completely random." He is talking about the whole nature of picking candies from a well-mixed jar, that is the point of reference for "it's so random." Thus, possibilities and likelihood, proportional reasoning, and variation come together for JM to constitute his explanations for what he expects.

Displaying Variation

JM drew significant attention to the averages, but also commented on shape and distribution in his reasoning about the graphs.

Evaluating or Comparing Graphs: While the other cases profiled so far also attended to the averages, JM repeatedly made the average a focus of his commentary. Consider the explicitness he offers in talking about Q8 ("35 Muffins"):

> JM: Well, I look at the median as it's written out {Summary box}, and I also look at the amount of muffins at 113 {Histogram}, which is… you know, the mode is actually 113 too, and here {Boxplot} it's also the median …When you look at the histogram, right away, you know, it pops out: Boom, 113. Even though it's uh… It shows it there {Boxplot} quite well too, in reading that

Although he misreads the median from the boxplot as 113 g (it is actually 113.5 g and is also given in the summary box of statistics), he moves his focus from the averages in the summary box to the mode in the histogram and then to the median in the boxplot. He also compares means for the two bakeries in Q9, a strategy not made as explicit in the previous cases discussed:

> JM: The East has a high mode, it makes a bigger muffin, at uh… {142 g}, But the median is , is much higher in the West. So when you add up all the muffins – if you could add up all the muffins – You'd get more muffin dough with the West than with the East

It can be seen how JM compares the higher East End mode (142 g for the East versus 114.0 g for the West) to the higher West End median (112.5 g for the West versus 101.5 g for the East), and the incorrectly considers the total muffin weight in light of these averages. Later, he considers the higher mean, which is in the East End Bakery:

> JM: It's higher in the East than it is in the West. Not by much, you know, half-a-gram… which, um, when we do the typical or the average, and we look at the mean here, we really wouldn't want to look for the typical muffin in the mean, we'd want to look for it probably more in the median.

This is a significant statement on JM's part, because it shows an appreciation for one of the concepts to come out of the Math 212 curriculum, namely the appropriateness of different measures of center for different situations. In the situation presented by Q9, JM made a reasonable discernment about the variability in East End Bakery, saying that "the range is too many, uh, low-weight and high- weight muffins. That really throws off our idea of the typical muffin."

Besides the averages, JM also mentioned the use of ranges and distributions. With the "35 Muffins" of Q8, for instance, he identified the maximum and minimum from the boxplot and also the IQR, saying

> JM: And when you look at the box and whisker plot, we see, uh… Well, the range of course is what – between 112.5 or… and 115.5… and 50% of them fall within this range

JM's latter use of the term "range" was in reference to the IQR, which I could tell because his prior comments had to do with the entire range; he used the same term for range and IQR, but knew that the IQR held about the middle 50% of the data. The distribution of data figured into JM's reasoning on Q9 ("Two Bakeries") as JM expressed concern over the lack of any muffins weighing in a span of values for the East End Bakery: "You would look down here {East End}, and of course you can see this big gap between 102 or something and 130- something, that there are no muffins." It should be mentioned that East End's muffin weights are split into two groups roughly clustered towards the ends of the distribution with the approximate gap to which JM did take note. JM also reasoned about range in Q9 when comparing the averages of the two bakeries, and he even compared spreads of data shown in the two boxplots:

> JM: And the range in the East End bakery is tremendous, really… The interquartile range – 50% of all the muffins – are just , uh… exceeds the whole range of the West End bakery

This is a powerful observation that JM makes, and he uses it later in his making a conclusion about the bakeries, which will be discussed in the next dimension of this aspect of displaying variation.

Outside of the muffin questions of Q8 and Q9, JM also paid attention to distribution in the sampling context of Q5 ("Small & Large"). He called attention to how, "for both of them, they're both the same. They're both higher in the 50 percentile range. 50% are higher…" Since frequency graphs using

bars were given in both of the graphs in Q5, I needed clarification on what JM was talking about with the "50 percentile range" and the reference to "50% are higher." It turns out he was noticing how the data in both graphs looked like a major portion of the data – possibly 50% if drawn as a boxplot – was clustered near the expected value (6 reds for the Small Jar and 60 reds for the Large Jar). He also question the amount of data distributed at the higher and lower ends of the distribution. Finally, in the probability context of Q12 ("Compare Comments"), JM felt that even though the data as shown was possibly from an unbiased spinner, the distribution would look better if it had a different shape. He suggested the following idea:

> JM: I think the distribution of the 20 {trials} – of the 50 spins {per trial} would be more symmetrical, with 25 probably being the mode, and probably being the median, and just falling off…{He has drawn an inverted V with mode at 25 on the graph}

By "falling off," JM meant a tapering of data away from the mode, as shown in his inverted "V" shape. It was interesting how he had drawn the same kind of shape on the PreSurvey, and on another written instrument where a bell-shaped curve or skewed bell was called for. Regarding levels of detail and subsequent usefulness of the different graph types, JM showed a preference in Q8 ("35 Muffins") for the way that the histogram showed the mode, saying that "the histogram is {a} really easy, graphic display for just about anyone to see, it's 113 is the one that shows up quite often." He explicitly commented on how the boxplot had less detail while the histogram had more:

> JM: Well, I can see from the boxplot that the low point is 109, it

> doesn't tell me how many, of course, that's one thing. Whereas when I look at the histogram, I, you know, I can see every muffin just about, and how much it weighed. And if I was really concerned with each muffin, I'd know from that {Histogram} really well.

Conversely, in Q9 ("Two Bakeries"), JM shows a preference for the coarser presentation given in the boxplot, and says

> JM: Here, I like the box-and-whisker… Maybe because it's bigger, and it's just – I like the way they laid it out, and it helps me better. I mean, you know, you could get the same from this {histogram}, I think, but…I like the box-and-whisker here, because I could easily see the range

It seems from JM's earlier response whereby he compared the IQR of the West to the entire range of the East that boxplots were instrumental in his observation and subsequent conclusion, which will next be shared.

Making Conclusions: For the Q9 ("Two Bakeries") situation, it is interesting that JM's strongest conclusion statement actually came at the beginning of his commentary:

> JM: Well, obviously the West End bakery produces a , on average, a bigger muffin – whether or not it's better, we don't know – but it's bigger: It weighs more.

Later in JM's analysis, it became apparent that the East End had a bigger mode and a (slightly) bigger mean, and only in the median value did the West exceed the East. It seems that the median, as shown in the boxplot and summary statistics, is what instigated JM's first judgment, plus the fact that the variation in the West was smaller than in the East. Even though the largest muffins are produced at the East, JM felt that he might not get those higher weights:

JM: So, I'd probably be less confident going to the East End Bakery, unless I knew on which day they were producing the big muffin, if I could find out… If that's what I was after

His other conclusions about graphs were directed at Q12 ("Compare Comments"), where he thought the distribution was acceptable (and also offered what he felt was a better shape for the graph), and also at Q5 ("Small and Large"). In the latter question involving the comparison of the two graphs, after voicing initial suspicion over how similar the graphs looked, he concluded that "maybe it's not that surprising." Part of his rationale included how such results were possible, and when asked if he thought one graph or the other was likelier to reflect made- up data, JM said "I don't know, it's hard to say." I re-phrased the question in terms of whether one graph's results surprised him more than the other, and his response was: "Which one surprised me more? Oooh. I don't know, I don't know." I categorized this response as having not much confidence in declaring either graph real or made-up, which is basically making *no* conclusion.

Interpreting Variation

JM had more to say about causes of variation in the PreInterview than in the PostInterview, but he still did offer some idea about his thinking in that dimension, as he also did about some effects of variation. Most of his responses were directed to influencing expectation and variation.

Causes of Variation: A couple of JM's comments in the sampling and probability contexts suggest that there is a link in his mind between

randomness, luck, and variation. Earlier he had referred to drawing from the Large Jar as “completely random”, and in Q1c (“Pick Six”) he talked about how he’d “just feel lucky with that handful, with those handfuls.” For many people, chance, randomness, and luck are among the ideas associated with the presence of variation, and it seems that JM holds similar connections. Note how on the PreSurvey, JM wrote a definition for random as “chance or arbitrary,” and for the other device involving uncertainty – the spinner – he had suggested for Q10a (“One Trial”) that there was going to be “some variation” involved. It is possible that JM sees luck or randomness as a cause of variation in the sampling and probability situations.

There was also a physical component to the spinner which JM commented on as a source of variation. In Q10b (“Compare Trials”), he said

> JM: Well, you know, I was thinking, it is a fair spinner, right…No matter where he starts the spin… You know, that could have some, you know, you start getting into the rhythm of it… And, affecting it somewhat. But if it’s a really good spinner, that’s… You know, it’s gonna be close to the first ones…

Although it isn’t explicitly stated, it seems that JM is relying on personal experience in his pointing to these physical causes of variation with the spinner. In class we had done some work with spinners, and people did share the opinion that if the spinner was launched from the same initial position and with the same force, then it could be expected to land in about the same region each time. In fact, the “rhythm” can be seen as a pattern of behavior emerges in spinning: Launch the spinner, record the results, then “re-set” the

spinner (to the same initial conditions), and then repeat the action. This behavior would be akin to holding and rolling the die the same way each time. The view is then held that if you don't roll or spin the device randomly, you can't be expected to get random results. Hence, it is how one rolls or spins that causes the variation.

Effects of Variation: One of the effects mentioned in the "Two Bakeries" scenario of Q9 was that JM had less confidence going to the bakery with greater variation in muffin weights (the East End) bakery, because he would not know whether or not he could get one of the bigger muffins produced. He also said that he wouldn't even be confident of "getting the average muffin," suggesting that "I might get a small dinky one or a big one." The common theme is that variation impedes one ability to know for sure what will result, and this idea came out further when talking about the sampling context. In Q1a ("One Trial"), he mentioned that "don't know because it's all random," and in Q1b he said for Repeated Trials" that "I think it's a good chance of it being within that range, whether it makes it or not, who knows?" An interesting comment came outside of the common subset, on Q2 ("Lists" for the Large Jar) when JM opined that "One man's range is another man's …randomness." Rather than whimsy, this comment does appeal to an effect of variation which might be otherwise expressed as saying that although results might be expected to collectively fall within a particular range with a certain degree of certainty, individual results still fall randomly within that range. Uncertainty is a core effect of variation, and this shows in JM's

indecision about Q5 (“Small and Large”), where he repeats “I don’t know” when considering the validity of the two graphs.

Influencing Expectation and Variation: JM had several ideas in this dimension involving the influence of the number of pieces used in sampling situations and the number of trials performed in probability and sampling contexts. His first main insight was shared in Q5 (“Small & Large”), when he was mulling the similarities between the two graphs. He had already been talking about how it was possible for the graphs to be similar in shape:

> JM: Right. Well, I think it’s, I think it’s possible, but I think you get a – Just, with more numbers, um, the larger container, you get a better idea of , of the actual number in the container.

When he talks about the “actual number” in the container, this could either be a reference to the actual mix of candies or the ratio of red to yellow. In class when we had done a sampling investigation called “Mystery Mix,” we had teams of students looking to predict the numbers of candies or to give the mix – that is, to state the ratio. JM also talks about “more numbers {in} the larger container” and possibly is thinking of both the higher number of candies in the Large Jar and the concomitant higher number of candies in a handful from the Large Jar. He is certainly explicit about a link between sample or population size and the power for inference, even if he is not quite clear about how he sees that link. JM is more explicit about how “the graphs would – They would look the same, if they did enough pulls, enough sampling,” and this ties the number of trials in with the overall shape of the graph, his second insight. He continues the theme of number of trials and the shape of the

distribution when discussing Q12 (“Compare Comments”), saying that after 20 trials the distribution of results should be fairly symmetrical with a mode and median at 25 blacks. JM added: “So, you have enough spins, it’s gonna look good!” (referring to his inverted-“V” shaped graph that he’s drawn, with mode at 25 blacks). He had also mentioned increased number of spinner trials in connection with actually attaining the expected value, as in Q10a (“One Trial”) when he said “…and if he does it enough times, he’s going to be right at that number.” Extending this notion into his final insight, he links the number of trials to the cumulative average of results, as we had done in class, and for the spinner trials JM gave 50 trials as “a fair amount, that’s a fairly good sampling, or trial run… that would approximate the theoretical probability.” Thus, JM suggests the ideas that in sampling contexts, sample or population size affects results that could be used for making inferences, and also that the number of trials in a sampling or probability context affects expectation, variation, and the shape of the distribution of results.

Discussion

There are some ways in which JM showed a shift in emphasis in his responses moving from the PreInterview to the PostInterview, and some other ways in which he showed some stability. One shift has to do with *what* is expected. Whereas JM put responses like “maybe 6” for PreInterview Q1a (“One Trial” at the Small Jar) and all tens for PreInterview Q9 (“Sixty Tosses” of the die), his PostInterview responses for expectation all had some form of a range. A more important shift has to do with JM’s consistency in terms of

what is possible or likely, a part of the *why* dimension for the aspect of expecting variation. In the PreInterview, JM showed some major inconsistency in terms of what he regarded as possible. Within about three minutes of the PreInterview, on Q1b ("Repeated Trials"), we had the following exchange:

DC: Ok. So, you put the handful back in, you mix it up, you pull out another handful. Is it going to be that every time?
JM: No, no. Of course not.
DC: You seem pretty strong about that. Why not?
JM: Well, because it's… it's impossible.

JM was in fact very strong about the impossibility of repeated results looking the same each time, and what is interesting is how, for so many other instances in the PreInterview, there were qualifiers added to JM's thoughts about how outcomes were possible. In fact, contrast the above comment for the sampling context with JM's later comment for repeated trials with the die on PreInterview Q11a ("Repeated Trials"). Recall in that question how I had subjects look back at PreInterview Q9 ("Sixty Tosses") and look at how many "5s" they had predicted in sixty tosses. JM had put all tens for each face of the die, meaning he had predicted a result of ten "5s".

DC: Are you going to get that many 5's again?
JM: Maybe
DC: Can you talk a little more about that? 'Maybe.' Would you be surprised if it was the same?
JM: No, I would not BE surprised.
DC: So you got ten the first day, remember, because you predicted that, and the next day is ten again. That's alright with you?
JM: It's ok.

Although there are phrasing differences in terms of getting the same results "every time" or "again", the sense comes through very well that by the end of the PreInterview nothing is as impossible for JM as he indicated at the outset of the interview.

Furthermore, in PreInterview Q1c ("Pick Six"), he noted that "you know, of course, you CAN pick out ten red, or you can pick out zero red.," suggesting that extreme values were possible. But by the time of PreInterview Q4 ("Real: 300"), in looking at the purported results from 300 trials at the Small Jar (which were in fact genuine), JM made the following series of comments throughout his analysis:

JM: I don't think you COULD pick up all 10 red, or all…zero red
JM: I just think it's impossible to pick out {all reds} if they're mixed
JM: I don't think I would've included all ten, because I think it's impossible to pick up, in your hand, a handful with no red.
JM: I don't think it's possible to get a handful of either all red or one red.

I included the whole series of comments because this represents a major change of expectation for JM. By the time of the PostInterview, after activities had been done in class, JM was no longer mentioning extreme results as impossible, but speaking more in terms of being unlikely. The shift in emphasis seemed to go from possible- or-impossible to possible-but-unlikely.

Some stability was shown in JM's ability to reason about displays of variation in the sense that he showed a fluency in using the graphs to tell him about the different centers and spreads of data and what the consequences of those differences would be on both Pre and PostInterview. For instance, in

PreInterview Q8 (“MAX Wait-Times”), he started off with a focus on centers, saying that “looking at the graphs, they’re both the same, right?” While saying this, he was pointing to the boxes of summary statistics, which were identical for both data sets portraying the Westbound and Eastbound trains. Later in his analysis, he determined that the train with smaller variation (the Eastbound in this scenario) was “much more reliable,” and made comments about the likelihood of differing wait-times depending on whether one was using the Westbound or Eastbound train line. He used very similar strategies in the PostInterview Q9 (“Two Bakeries”) situation, again first reasoning about centers and then also taking into consideration spread. Because the boxplot representation had been introduced in the Math 212 class, it was interesting to see how facile JM was in using the boxplot as he did to compare the IQR of the East End Bakery to the (shorter) range of the West End Bakery. It was especially noteworthy how JM knew that a coarser comparison was appropriate in some ways for the “Two Bakeries” scenario of Q9, while a more detailed graph such as offered by the histogram was more appropriate in Q8 (“35 Muffins”) for making a prediction about how much his own muffin might weigh. Knowing which graphs might be useful for which purposes in displaying variation is at the higher end of reasoning in this aspect, and it seemed that JM had some consistently adept ways of handling the questions involving graphs in both the PreInterview and the PostInterview.

There was another interesting comparison to be made in the way JM volunteered causes of variation in the PreInterview and the PostInterview. JM

had many more ideas that he suggested in the PreInterview than in the PostInterview. For instance, in the PreInterview Q1c ("Pick Six" at the Small Jar) he suggested that one might get different results "depending on how they're mixed up." He seemed especially interested in how the individual candies might lie in the jar next to one another, and in PreInterview Q3 ("Fake: 30") – involving thirty trials at the Small Jar – he gave expressed his thinking as follows:

> JM: Uhhh, <Long pause>. You know, I guess I'd have to see how they fit into your hand. <Chuckles> , Maybe that has a bearing on it possibly, right? And when you reach into a container, and pull them out, and if they're completely mixed, whereas one red is lying is against, or there'd be, what is it, 60%? So you'd have almost …2 reds around 1 white {yellow}. Maybe? Something like that?

So too did JM stress causes in the spinner scenario of the PreInterview Q12 ("Surprise Spinner"). I quote the entire exchange because it really shows JM's emphasis on physical causes of variation:

> JM: Um, well…I want to look at the engineering of the spinner, where do you start the spin, you know, I mean…. Do you start it in white, you know, the velocity, or the force… None of that really matters, I guess…I mean, it CAN matter of course, yeah. Well, of course, it WOULD matter, you know, I mean, you play like a game that has a spinner, and , if you're a kid, you know if you hit it just the right way, and you start it at just the right the spot, you could… there's a chance of it being in one spot are greater than in another spot.
> DC: So this is very well-oiled spinner…Very, very fair spinner
> JM: Ok, so this is a GOOD spinner. Yeah. Ok. A fair spinner. Um, yeah. And the spinner is, is flat? A flat plane? It's a fairly spun game?

Rather than being contentious, my sense was that JM's expectation of variation depended greatly on the physical apparatus and the actual performance of each trial, whether it was drawing candies from jars or using spinners. However, in the PostInterview he offered very few ideas about causes in these contexts, and it seemed that his aside comment above about "none of that really matters" probably gained dominance over his thinking as we engaged in the class activities designed to show random behavior. He did mention the way the spinner was used in PostInterview Q10b, and "getting into the rhythm of it" - as was shared earlier in the case study. However, there was much less in the way of JM volunteering such physical causes in the PostInterview than in the PreInterview.

Finally, while JM did have some references to the dimension of *influencing expectation and variation* in the PreInterview, he had many more references in the PostInterview. An example of the kind of thinking he had in the PreInterview came from Q12 ("Surprise Spinner"), when he noted that

> JM: Since two-thirds of it is white, it seems that , over the long run, two- thirds are going to come out white, over black. Now, you might not do it in 3 spins, you might not even do it in 12 spins, but if you do it in 10,000 spins, I think you're going to be closer to 66, 660 whites , you know, or whatever…

He is appealing to the Law of Large numbers in the above example, and in the PostInterview he maintained his earlier opinions in this regard, while also adding more references to the expanding range expected as result of performing more trials. Some examples of JM's interpretations from the common subset of PostInterview questions were shared earlier, but there

were many more responses from outside of the common subset. For example, in the sampling context, in PostInterview Q4 (“Real: 300”) JM said

> JM: I think it’s just in 300 pulls, I mean, it’s gonna happen, you’re going to pull out less than 48 reds, at least once.At LEAST once. Maybe twice, or three times, or four or something

He had a similar thought for the spinners in the probability context, about attaining more extreme values with increasing numbers of trials. He also articulated again his interpretation of the Law of Large Numbers in PostInterview Q13 (“Real or Fake?”), a question which showed purported results from two classes who were to have each done 30 trials at the spinner:

> JM: So, the theoretical should come close to the experimental… over the long run, if we do enough trials, and have a big enough sampling of what we’re doing. So once we figure out the theoretical, we go out and try to prove it experimentally, and see how close they come. And , chance are, they’ll come pretty close if we do a fair number of sets.

His ideas reflect well what had been done in class.

In summary, JM showed some stability in some regards, such as his fluency with graphs, and some shifts in others, such as his concern over physical causes or his sense of what was possible or likely. Overall, while still exhibiting responses showing the strong influence of the average in this thinking, he had some reasonable ideas about expecting, displaying, and interpreting variation. His ideas seemed to grow in maturity as a result of class experience. Particularly with respect to influencing expectation and

variation, he had a fairly well-developed notion of what an increase in number of trials might do to results in a sampling or probability environment.

The Case of SP

The first interview showed me that SP was a very reflective individual, someone who really thought not only about her answers, but also how she was thinking and feeling about the questions. Her language suggested she was comfortable with a sort of metacognition, and she repeatedly talked about her instincts and feelings, often contrasting those thoughts with a logical perspective. For example, she would talk about "my first instinct", and then how "there's not any super-logical reason" but "I guess that's just where my brain goes first" – She clearly showed a willingness to try and explain what was going on in her mind. She volunteered information readily - for instance, telling me what would or wouldn't surprise her – and was an easy person to talk with; however, both of her interviews lasted a bit shorter than average.

SP had taken Math 211 the prior quarter with Steve, and wrote on her PreSurvey that she had taken some probability or statistics course at another university four years ago. She recalled that it had been a "fun, interesting class," yet currently she said she felt "comfortable but shaky – don't remember much but I'm sure it will come back to me." She wrote that that variation meant to her "the differences between things in a group," and gave several examples: "Weight, height, hair color of a group of people."

There is a single theme of SP's that was so prevalent on the PreSurvey and PreInterview that it was reflected in several ways which I'll highlight before discussing her thinking on the PostInterview questions. The theme was how "Anything is Possible", and although all the other cases pointed out the possibilities for different results, no one among the six cases repeated this theme more than SP. On the PreSurvey, in looking at "One Trial" with flipping the coin, she wrote that "It could be any number of times," and for "Six Trials" she had included a choice of 2 heads out of fifty flips, saying she "Just chose randomly – Anything is possible." Then, for her very first response on the PreInterview, concerning one trial at the Small Jar, she wrote: "My first instinct is just to say that it could be any amount."

The responses shared so far are only a small sample of how strongly this point of view came out in SP's PreSurvey and PreInterview data, and the theme of "Anything is Possible" is related to two other themes I noticed: Difficulty in making choices, and the sense that one can never know for sure what results will be. For SP, it was often difficult to make decisions about what was or was not real data, and it was also sometimes difficult for her to even venture a guess about results. On the very first question for the second part of the PreSurvey, "One Trial" at the Small Jar, SP was the only person to *not* give any number at all (such as the common "6 Reds" answer). Instead, SP wrote "Hard to say." In looking at the graphs on the PreInterview to decide if sampling results were real or fake, SP said she was "attracted to [the opinion] 'We have no confidence,' because we REALLY can never know." Later, in

looking at the dice questions on the PreInterview, she said “I guess…I can never know.” The themes shared above were hardly reflected in the PostInterview, a contrast which will be considered again after the PostInterview results have been discussed.

Expecting Variation

In this aspect, SP had noticeably more to say about *what* she expected as opposed to *why*, but she did have reasonable responses in both dimensions.

What was Expected: Whereas the theme of how results should be close to the expected value or average came out very explicitly with the other students, SP hardly spoke of results being close to the expected value or average using explicit language: That is, she rarely said in the PostInterview that she expected to get around 6 reds, for example, or close to 25 black. Rather, she talked about two other themes within this dimension of *what* was expected: One theme concerned how results should not repeat as often as they should be different, and the second theme concerned the range, with the latter theme being the most frequent way that SP had to talk about *what* she expected. SP did refer to the expected value, but most often it was in the context of results averaging out to or being within a range around that value. For example, in Q1a (“One Trial” at the Large Jar), her first response on the PostInterview was to suggest a range, “somewhere between 50 and 70,” and she also gave a range on Q10a (“One Trial” at the spinner). She went on in

Q1a to suggest that “you expect sort of an average of 60, if you did many of these,” and in Q5 (“Small & Large”) she noted that results were “going to be near to 60.” The latter response was as explicit as SP ever got as far as having expectations that directly concerned results being close to the expected value or average.

Instead, SP had comments and responses showing how results should not repeat every time. In thinking about several trials in Q1b, or comparing two trials in Q10b, SP emphasized how she would expect “not the same number” (Q1b) but “just different numbers” (Q10b). Her own choices for six trials at the Large Jar and at the spinner showed no repeated values, and she was even more articulate about the theme concerning repeated values in two of the questions that aren’t a part of the common subset. The two questions were Q2 (“Lists” for the Large Jar) and Q11 (“Lists” for the spinner), and SP repeatedly judged lists in part on the basis of how many repeated values she saw in the list. Here is a sample of her responses:

SP: (Q2) Each time you choose, you’re most likely going to get a different result rather than the same result, over and over again.
SP: (Q11) Again, you wouldn’t expect to get the same exact thing, I expect more variation.
SP: (Q11) You’d expect there to be greater variation and less repetition

Clearly, SP does not expect to see many repeated values, and in Q11 she makes a connection between more variation and less repetition, which fits very well with her thought on the PreSurvey about the meaning of variation having to do with the presence of differences. SP does allow for the possibility

of repeated values, as she noted for “Six Trials” in Q10c (she had put “20, 23, 24, 26, 28, 29”) : “They could repeat, but I just did a range from 20 to 30, just to choose.”

The focus on range or extremes was the second theme apart from the theme of repeated values, and as mentioned earlier, this theme was the main way SP used of expressing what she expected. In Q1b, she expected repeated trials to be “somewhere in that range,” meaning the range of 50-70 which she had put for Q1a and within which all of her choices for six trials in Q1c fell. She used the same exact words in Q10b (“Compare Trials”) as she had in Q1b, and reinforced her opinion by adding: “I think the range would still be somewhat similar.” When evaluating arguments for the spinner results in Q12 (“Compare Comments”), SP was clear that as long as results were “within that sort of 20 to 30 range,” she would wouldn’t be to surprised how the data was distributed within her range of expectation. In fact, SP considered a reverse situation, saying “it’d be surprising if there were NONE between 20 and 30.” The main point here is that SP continually talked about her expectations not so much in terms of a single value as much as in terms of ranges and extremes. In the “Small and Large” scenario of Q5, she narrowed the ranges shown on both graphs, saying that the more limited range (with the lower extremes crossed out by SP) was more what she would have expected. Finally, even in the situation of Q8 (“35 Muffins”), when asked how much she expected her muffin to weigh, SP also gave a range answer: “I would expect it to be somewhere between, like, 112.5 or something to 115.5.”

It didn't surprise me that SP, who had focused so heavily on how anything was possible in the PreInterview, would be attracted to the two themes of how often results might repeat and how a range was expected for results as opposed to a single value.

Why (Reasons for Expectation): On the common subset of PostInterview questions, some of SP's responses reflected the theme of possibilities and likelihood, but she used a different kind of language to get across the same idea of the theme. That is, she talked about how she "might expect" a certain result, or what "could happen," but she didn't talk as much in terms of how anything was possible as she had done in the PreInterview. She also seemed to use the language of what was "surprising" as way of talking about what was unlikely. For example, in Q1b ("Several Trials") SP said that eventually she would expect "some more extreme numbers," tacitly acknowledging the possibilities of extremes. On Q1c ("Six Trials"), in explaining *why* she had chosen "49, 51, 55, 62, 65, 68," SP's response shows the theme of possibilities and likelihoods (even though she doesn't use those exact words):

> SP: I guess I went a little crazy with the 51 and 55, but you might expect to see that too, every once in a while have one that's more on the outskirts – Well, 55 you would expect, but 51 would be "maybe" , or 49, would be on the "more surprising" scale.

There could be some disagreement about how surprising 49 really is for six trials, which is really a part of the whole theme of speaking in terms of what's possible or likely – That is, there is some subjectivity behind what could or

might happen. Similarly, on Q12 (“Compare Comments”) SP first noted “it’s not THAT surprising that someone didn’t get exactly 25,” and again the way she uses her language suggests that SP is telling what she thinks is *likely* or *unlikely* when she speaks of what is or is not surprising. For the results of twenty trials at the spinner, SP is effectively saying that it’s not unlikely for the expected value to 25 blacks to be missing from the set of results. In considering the upper extreme of 34 blacks, she said:

> SP: Yeah, it is a little higher than you would expect, but …Definitely could happen.Yeah, the 34 is surprising, but there was only one of them, so it’s fine to have that bit of surprise.

An example of SP specifically using the language of possibilities or likelihoods came from Q2 (“Lists” for the Large Jar), when she was talking about how results should not keep repeating, saying “this is very unlikely, that you would pull the same number, again and again and again.”

The lack of emphasis on the expected value in the sampling and probability contexts made me curious about whether or not SP was actually capable of reasoning proportionally. I did find evidence that SP knew how to calculate ratios, and so I suspect that it is not the case that SP cannot reason proportionally, it is instead the case that SP is not overly influenced by the proportion, and more will be said about SP’s reasoning in the discussion. An example of SP reasoning proportionally comes as she consider the two graphs in Q5 (“Small & Large”):

> SP: (Q5) Well, because the ratios are the same. Oh wait, it’s a container of 100, and a container of 1000, okay. Yeah. So, the

choosing 10 out of 100 is same as the ratio to 100 out of 1000

Also, since she so often talked about ranges and so seldom mentioned the expected value, occasionally I probed to see *why* she gave a range, as in Q10a ("One Trial" at the spinner):

> DC Oh, between 20 and 30. Well, why do you think that?
> SP Umm, because of that 50 to 50 ratio, or chance of getting black, and chance of getting white – And so, out of 50 times, half of 50 is 25, and so that would be the, sort of – expected ratio. Not expected, but the – Theoretical ratio {Laughs}

Her language at the end of her response shows much about *what* SP expects and *why*: The expected value (from the theoretical ratio) is not really expected for SP, and instead she expected a range around that value (of 25 black). Despite her lack of emphasis on the theoretical ratio, SP definitely could reason proportionally.

The theme of variation as an explanation *why* SP had her expectations only came through explicitly as she considered the spinner, expressing that physical causes of variation made her think of putting a range answer. More will be said about the physical causes when talking about SP's *interpretation* of variation. Implicitly, SP seemed to be relying on the notion of variation as corresponding to differences in results (akin to what she had said about variation on the PreSurvey). That is, she continually stressed ranges for her expectation, and she also said that results should look different, and so I suspect that SP has a dual notion of variation as both an expectation and a reason. In saying that she expects a range because results will be different, essentially she is invoking variation as a reason *why* a range is expected. On

Q12 (“Compare Comments”), when offering reasons for supporting or disagreeing with the provided arguments, SP also used a bit of variation in connection with her distributional reasoning, pointing out how data could be clustered and still be plausible. For example, she implicitly acknowledged that the median and mode could vary from 25 blacks in twenty trials at the spinner. More on her reasoning about the distribution follows in discussing the aspect of *displaying* variation.

Displaying Variation

Just as SP had little focus on the expected value in the sampling and probability contexts, so too did she seem to pay minimal attention to the average when considering displays of data. Instead, more of her responses had to do with spread, shape, and the conclusions she made concerning the graphs.

Evaluating or Comparing Graphs: SP knew where the averages were on the graphs, and she knew what they meant, but my basis for this claim about SP’s knowledge comes more from her reasoning in class and on the survey instruments than from explicit responses on the PostInterview. That is, whereas some of the six cases repeatedly used the words “median”, “mode”, or “average”, SP just didn’t invoke the terminology to show any sort of focus on the average in her responses. She physically pointed out the modes when comparing the graphs in Q5 (“Small & Large”), saying:

> SP: And so, you would expect, sort of, the results to be somewhat similar, but just on the 6, it’s going to be near 60. {SP points from

the mode of 6 on the Small Jar graph to the 60 on the Large Jar graph}

Notice how SP calls out the values “6” and “60” but does not name them as modes. Her style is similar in Q12 (“Compare Comments”), where she refers to the expected value of 25 but never uses the term “average”.

Instead of the average, SP’s responses had more to do with the spread and shape of the distribution. For example, in looking at the graphs for 35 Muffins in Q8, SP voluntarily circled a grouping of data on both graphs as her immediate response:

SP: How much would I expect my muffin to weigh? Well, I’m guessing that it could be anywhere in between, somewhere around where the bulk of this data is, {Circling some data on the histogram and then a region of the boxplot} probably … So I would expect it to be somewhere between , like, 112.5 or something to 115.5

Her values of 112.5 and 115.5 grams correspond to the first and third quartiles on the boxplot, so SP was focusing on the interquartile range. Again, as an unprompted response, when I asked for any similarities and differences in the way the two types of graphs showed the data, SP said that the histogram “shows you the greater variation,” and notice how the meaning of variation in her response does not equate to range, but to the differences among the set of data (consistent with her earlier sense of the meaning of variation). When comparing the two bakeries on Q9, SP looked not only at entire ranges, but again at the middle 50% of the data:

SP: The West End bakery is more consistent, there’s less variation, you can see that from here {Pointing to Dotplot} it goes from there to there, whereas this one goes boop-ba-doo {Marking the

> min and max on the dotplots for both bakeries}, and it's more spread out, and you can see that very well up here too, {Pointing to the Boxplot} too, that the center half of the data is much more spread out… Which you also see from here {Dotplots}

In the above response, SP really shows her strength in reasoning from both types of graphs, with a focus on variation rather than averages. There is a shift in the meaning of variation as she refers to the West End bakery as having "less variation" and then supports her reasoning by pointing out the shorter range. In Q8, when she had said that the histogram showed her "greater variation", she was using the terms to mean a muffin-to-muffin comparison of differences rather than the entire range.

As a final example of SP's reasoning about graphs along the themes of range and distribution, consider her remarks about Q12 ("Compare Comments"), when she said: "As long as there's – like, the bulk of the data falls in between that range [20 to 30], then I would think that it's fine." SP used that same term, the "bulk" of the data, in Q8 when talking about the 35 muffins, and to me it shows a focus on the clustering of data. She went on in Q12 to suggest that she would actually expect the "bulk" of the data to be shifted to the left a little, and she showed me what she meant by drawing on the graph:

> SP Instead of the bulk being here {She circles the data from 28 to 31} I would want the bulk to be there {She makes another circle around 24 to 26} But I would think that if would just do it over and over again, more than 20 sets, if you did 400 sets, whatever, that you.. that that would end up leveling itself out…

I wasn't sure what she meant by "leveling itself out," and so I asked her further about that idea, and she drew a bell-shaped curve over the data. It became

apparent to me that SP had some good reasoning abilities about displays of variation. Although she didn't use the same terminology or focus on the average, she had sense for spread, variation within the data, and the shape of the distribution. Also, she was comfortable using different types of graphs in making her comparisons.

Making Conclusions about Graphs: Most of the responses already shared above show the conclusions that SP made about the different graphs. For instance, in Q9 ("Two Bakeries"), SP noted that the West End bakery was "more consistent" because the range of muffin weights was less than at the East End. She also said that both types of graphs (dotplots and boxplots) were equally useful to her in making decisions. Also along the theme of usefulness and the level of detail provided by different graph types, SP had noted in Q8 ("35 Muffins") how the histogram told her more about the variation among the data. She went on to explain how the histogram gave her more detail than the boxplot:

> SP: Because {on Histogram} you're getting each individual number, along this line, whereas this {the Boxplot} is just showing where the center half of the data is, and then, where it begins, where it ends…

For the boxplot, SP further explained, "you're not really getting any levels of how much is there, you're just getting that there WAS one there." She showed me how, for instance, there might be some data within the whiskers of a boxplot, and there would be data at the maximum and minimum, but for "levels" she meant that she could not tell the frequencies from the boxplot.

In addition to conclusions about the graphs on Q8 and Q9, SP also had conclusions about the graphs in the sampling and probability contexts. In comparing the graphs on Q5 ("Small & Large"), she started off by saying "I don't think that you could, you could definitely suspect one or the other," meaning (I believe) that she could not conclude that one class was likelier to have made up the data. Later, she admitted the difficulty in making conclusions, saying:

> SP {Laughs} I would say that it's really hard – It's not that you would expect one more than the other of cheating, but maybe both of them, because these {circled groups of data on the low end of both graphs} are a little surprising?

So, although it was hard to point to one class over the other as having cheated, did conclude that the lower extremes seemed unlikely. Another conclusion she made was that, regardless of the veracity of the data, both graphs should look similar. Particularly if she imagined the graphs without the low extremes, she said: "Yeah, if I cross those [low extremes] out, that's what I would sort of expect, and you expect them to look similar, because it's the same ratio." With the graph on Q12 ("Compare Comments"), SP was clear in her conclusion that the graph looked like what she would expect for actual data coming from twenty trials at the spinner.

Interpreting Variation

For this aspect, SP had less to say about *causes* and *effects* than she did about *influencing expectation and variation.*

Causes of Variation: SP had several plausible causes of variation for the repeated-measurements muffin question of Q6 ("Causes: Muffins"), which wasn't a part of the common subset of questions for the other cases but may have been on her mind as she volunteered a reason for Q9 ("Two Bakeries"). In thinking about the differences between the East and West End's muffin weights, SP suggested that "they {West End} have this specific 'muffin-pourer' with this specific 'muffin-pourer cup', and they {East End} just do it, eyeball it…" I thought her comment about "eyeballing" the measurement fit well with her earlier comment on Q6, where SP suggested one of the causes of variation in the twenty 6th – graders' measurements was how "they have to sort of eyeball what they think the measurement looks like, what angle they're looking at the measurement thing from." Also, her comments on human error or differences in perspective on the measuring device reflected what had happened in the class activity during "Body Measurements," when many students were fairly casual in their measurement- taking.

Human activity was featured in her reasoning about the spinners too, and fits with the theme of a physical cause of variation. In Q10, she remarked that the range of 20 to 30 blacks "would take into account the actual practice of spinning it," which was her first allusion to the theme of physical causes. She expanded on this theme in Q11 ("Lists" for the spinner), when she was most explicit in explaining how "you'd expect there to be greater variation and less repetition because it's a human little activity." She actually detailed how the variation could be controlled:

> SP Because, there's – I guess there's variation in your own actions of spinning it, so maybe if you had, like, if you were spinning it, and you could use the exact same pressure every single time, or force, then you would expect it to land on the same thing over and over again… But if you are – because you're human, and you're going to do things a little differently, so you're going to come up with different numbers. Or something like that

Her argument is akin to showing how to influence expectation and variation, but I have listed her response here because it is such a good exemplar of how students really see the physical causes of variation in probability situations involving spinners.

Effects of Variation: As more will be said in the discussion, I'll just briefly mention the surprise I had in finding that the indirect effect of "You Can Never Know", so prevalent in SP's PreInterview responses, was missing from her PostInterview remarks. Also, aside from having a bit of difficulty adjudging the two graphs in Q5 ("Small & Large"), there were hardly any effects of variation noted by SP on the common subset of PostInterview questions. The idea on Q5 that one can't have much confidence is, I believe, not so much an indirect cause of variation for SP because her reasoning has to do with the way the graphs *should* look similar.

On a question not a part of the common subset, Q3 ("Real: 30" for the Large Jar), SP shows a direct effect of variation by commenting on how probability and reality do not always agree. She said: "It's random enough so that it's not like this perfect bell-curve, so it seems like more of a realistic situation because it's not perfect." SP shares the same language with DS in referring to the underlying distribution as "perfect."

Influencing Expectation and Variation: One theme that SP mentioned had to do with the number of candies in the jar, and in Q1a ("One Trial" at the Large Jar) she said she expected a range "but too many less [Reds than Yellows], because there's so many reds in there." In the PostInterview she mentioned this theme less times and with less emphasis than she had in the PreInterview.

The other theme that SP addressed in this dimension had to do with the number of trials performed, and she related the number to the average of results, the amount of variation expected, and the shape of the distribution. Even in Q1a, as she spoke of having more reds than yellow she mentioned expecting "an average of 60, if you did many of these [trials]." For the amount of variation, she had some ideas related to the sampling context that she shared in Q4 ("Real: 300"), saying that the range would be extended with more trials, "which you would expect with more pulls you would do, the more sort of outliers you would get, or the 'unexpecteds' you would get." Her idea in the sampling context that "the more pulls you do, the more often you'd expect to get that low chance" was balanced by a different contention in the probability context. In Q12, when thinking about the extreme value of 34 blacks, SP noted that " the more that you would do these sets, 50 spins, the more it would probably come back towards that 25 {Motions her hand from 34 back to 25}". I interpreted SP's response in Q12 to mean that more trials would give a cumulative average more towards the expected value. Finally, regarding the shape of the distribution, she earlier was quoted in Q12

expressing how a bell-shaped distribution would result from 400 trials, and she added: “When there’s only 20 sets, you expect there to sort of be, this, like, more random look to it.” The theme of more trials influencing the shape of the distribution had also come out in Q3 (“Real: 30”) and Q4 (“Real: 300”) in the sampling context:

> SP: (Q3) And you only did 30 pulls, so it’s going to look a little bit more scattered
> SP: (Q4) Yeah, that it would become more conformed to this perfect bell-curve, and that it would pull out a little bit more

Her latter response was in regards to doing increasingly more numbers of pulls, and shows how SP regards the numbers of trials as influencing not only the range, but also the shape of the distribution of results.

Discussion

In discussing SP’s conceptions of variation, I’ll highlight some key emphases she made, drawing attention to similarities and differences in her responses from the PreInterview and the PostInterview. I’ll also organize the discussion according the main aspects addressed by her responses.

When *expecting* variation, SP was consistent with her responses in several ways: She didn’t focus on the expected value, she frequently stated her expectations in terms of a range, and she emphasized that results should be different more often than they should repeat. Along with her lack of focus on the expected value went a minimal use of proportional reasoning. It is, of course, not possible to give examples of SP’s *lack* of emphasis on expected value and proportional reasoning. However, as I contrasted SP’s responses

with the other cases in both the PreInterview and the PostInterview, I could see that for both interviews SP just had much fewer responses where she explicitly talked about the expected value or used proportional reasoning in the sampling and probability contexts. An exceptional response from SP came in the PreInterview on Q9 ("Sixty Tosses" of the die), when she did list all 10s for the outcomes. She did not invoke proportional reasoning in her explanation, but instead talked about being "forced to make a guess" (an attitude which will be discussed later in the *interpreting* aspect). Later in the PreInterview, on Q11b ("Six Trials") when she had to predict how many 5s would result from repeated trials of sixty tosses, SP wrote "2, 5, 7, 10, 14, 20", saying "it's going to range." Also in the probability context on the PreInterview, when considering the outcome of sixty spins at the 2 White-to-1 Black spinner (Q12, "Surprise Spinner"), SP first response was in terms of a range: "Between 40 and 50 I guess, whites." In the PostInterview, range expectations continued to dominate her responses. The idea that results should not repeat very much was stressed in both interviews (again with the exception of the sixty tosses of the die, for which she later reconsidered). Here are some sample responses in both contexts of sampling and probability:

PreInterview

SP: (Q1c: "Six Trials" at the Small Jar) Yeah. I'd be more surprised if the same number kept showing up, as opposed to if it was just completely random

SP: (Q11a: "Repeat Trials" with the Die) I think that it'd be more random. It'd just… It could be, I think… I feel it would change every time.

PostInterview
SP: (Q2: “Lists” for the Large Jar) This is very unlikely, that you would pull the same number, again and again and again.
SP: (Q10b: “Compare Trials” at the Spinner) I mean, there’d be – just different numbers, but still somewhere in that range.

Although SP emphasized the differences in results, she also kept reiterating how it was possible that results could repeat, yet she believed this repetition to be unlikely.

The biggest shift in terms of SP’s expectations was that her choices were better in the PostInterview – Specifically, her ranges got much tighter. Also, although I believe SP reasoned proportionally in both interviews (based on few but pertinent examples), she seemed to center her choices more around the expected value in the PostInterview. Consider, for example, how on Q1a (“One Trial”) on the PreInterview, SP’s first idea was “I guess instinctually I would say that it’d be somewhere like in a median, like uh…4, 5…just instinctually.” Since most people said “6 reds” or “around 6”, and since SP had written on her PreSurvey for the same question “Hard to say” (and now she was saying 4 or 5), naturally I wondered if SP even knew that the expected value was 6 reds in a handful of 10. Subsequent questioning showed that SP did know the relevance of the ratio 6/10 = 60/100 to the Small Jar, but SP kept talking about expecting results in the “midrange” of 4, 5, or 6 – The midrange for SP was exactly (0 + 10) / 2 = 5, thus an expectation of 5 or close to 5 is what she seemed to be focusing on. For choices involving six trials, SP was not only wide in PreSurvey and PreInterview questions, but in

some cases she was extremely wide. On the PreSurvey, for instance, with six trials of flipping the coin fifty times per trial, SP's choices were "2, 3, 10, 22, 16, 25" for the number of heads. This list is low and wide, and the 2 and 3 are very rare outcomes. Similarly on the PreInterview, I shared on the previous page how she listed on Q11b ("Six Trials") the possibility of getting two 5s in sixty tosses of the die, again a rare event. In contrast, her choices for six trials at the Large Jar (Q1c) were "49, 51, 55, 62, 65, 68", and at the spinner (Q10c) she put "20, 23, 24, 26, 28, 29" : For both situations, her choices are markedly better than corresponding choices she had put prior to the class activities. Also, she generally gave range answers in the PostInterview, and her ranges were tighter and more likely than those given in the PreInterview.

Concerning the aspect of *displaying* variation, SP mainly showed consistency in her focus on range and distribution, and also in her minimalized verbalized attention to the average. That is, SP knew about the mean, median and mode (particularly in the PostInterview), but she did not specifically refer to these measures very much in either interview. Instead, on both interviews she talked compared and evaluated graphs mainly by talking about shape and spread. For example, in the PostInterview SP was quoted earlier comparing the two bakeries on Q9, talking about the spread of the data, and in Q12 ("Compare Comments") she repeatedly referred to where the "bulk" of the data fell. So too in the PreInterview (on analogous questions to the PostInterview), she spoke of spread and the how the data was "clustered" :

SP: (Q8 “MAX Wait-Time”) Eastbound train is more consistent. Because we have these clusters here {Points around 9 and 11}…You’re also not taking into account that there’s this spread {Shows with her hands}. That this {Westbound} has a greater spread compared to this {Eastbound}

SP (Q13 “Compare Groups”) Well, this one {Group A} has greater variation than this one {Group B}. It’s more spread out, it goes from the lowest is 13, and it goes up to 30. This one {Group B} is clustered within 17 and 23.

It seemed that SP reasoning in evaluating and comparing graphs was very stable from the Pre to the PostInterview. One difference that SP showed was in handling multiple graph types (since the PostInterview included more types of graphs than the PreInterview), and SP readily was able to reason appropriately off of boxplots and dotplots as well as histograms.

It was *interpreting* where SP showed the least stability and the most change in her emphases. The related three themes that were mentioned at the beginning of the case study – “Anything is Possible”, “You Can Never Know”, and “Difficulty in Making Choices” – were heavily subscribed to by SP in the PreInterview yet virtually absent in the PostInterview. A typical sort of response by SP in the PreInterview, much like the examples shared earlier, comes from considering several trials at the Small Jar (Q1b):

SP: I feel like it’d be somewhere in the same range. But it’s gonna be, it can be anything. Logically that’s what my brain is telling me, is it can be absolutely anything.

It was SP’s PreSurvey and PreInterview responses that got me thinking about how “Anything is Possible” , “You Can Never Know” , and “Difficulty in Making a Choice” could be seen as effects of variation (even if that is not

necessarily how the subjects see their own thinking). Historically, ideas such as SP's are linked to the Outcome Approach detailed earlier in Chapter 2. The essence of the Outcome Approach can be characterized by an attempt to look only at the next outcome of a probabilistic event, and transfers to the sampling context by focusing on the results of the next sample drawn. I do think that SP's earlier thinking was characteristic of the Outcome Approach, but the fundamental reason that any of that themes of uncertainty and indecision exist in subjects' minds is because of the "omniprescence of variability" (Cobb & Moore, 1997, p. 801).

This is why I have categorized themes such as "Anything is Possible", "Difficulty Making a Guess", and "You Can Never Know" as indirect effect of variation: It is because of the variation inherent in the situation that leads to uncertainty, and subjects don't fully understand how much or little results might vary. SP herself links her difficulty in making guesses to the idea that anything is possible at the beginning of the PreInterview, in talking about samples from the Small Jar:

> SP: I think I'm just pulling out a number because I'm feeling like I should make a guess. But I really don't want to make a guess. Yeah, because I feel like it really can be anything. And so making a guess is just like…. Just saying anything.

Even as she stressed how anything could happen with six trials at the Small Jar, she put her choices as "1, 2, 3, 4, 6, 8," and she said she wanted to "be just random about it [her choices]." When I questioned her about what she meant, she said "Well, I guess I just never want to say 0, and never want to

say 10." Her latter comment was right on the heels of having said that anything was possible, and this apparent conflict shows much about what is really behind SP's language and reasoning. Specifically, while she may "feel like it really can be anything," it seems that she innately appreciated the difference between what is *possible* and what is *likely*, a distinction that she articulated much better in the PostInterview. Even on the PreInterview question for six trials with the die, when she had put the improbably two 5s for sixty tosses as a choice, we had this exchange:

SP: Yeah. I'm not sure what the rules are, that I feel about this…But, I know it could be anything, but for some reason I still don't put the 1 or the 0, or the 60.
DC: Why not?
SP: I don't know {Laughs}. I really don't.

The contrast between SP's earlier emphases and her reasoning on the PostInterview is stark: Instead of anything being possible, reasonable ranges were likely. Instead of commenting how "you can never know," and expressing reluctance at guessing or having difficulty in making decisions, SP readily volunteered opinions which she then rationalized using more appropriate arguments than she had used earlier. Two other differences for SP were that she had less references to the numbers of candies in the jar on the sampling questions in the PostInterview, and at the same time she had more references to the way that the number of trials might influence expectation and variation.

In summary, it seemed clear to me that SP had moved from the idea that anything can happen to the notion that some outcomes were likelier than

others. The related theme of “You Can Never Know” gave way to a theme suggesting that while you may not know for sure about a given outcome, you can still make reasonable statements of expectation. Also, she had less difficulty in making choices and decisions on the PostInterviews, and her choices were more reasonable.

The Case of RL

Of all the students in Steve’s section, RL stood out in class discussions, on the research surveys, and in the interviews as having the most mathematically-oriented responses. As an example, when thinking about the chances of pulling ten yellow candies from the Small Jar in the PreInterview, RL began calculating aloud:

> RL: If there is a 0.4 chance of pulling yellow, and then there’s a 0.4 chance of pulling another yellow, then there’s a 0.16 chance of pulling two yellows. And if you’ve got ten, then you’ve got 0.4 to the tenth, which makes it real unlikely

Aside from the way that RL was considering drawing a candy and then replacing it (as opposed to the intent of the sampling scenario, which was to pull the handful without replacing any of the ten candies), his response is unique in that no one of the other cases showed such a willingness to calculate to the same extent as RL. He had a strong background in mathematics and also in philosophy, and during the interviews he would occasionally veer off on some tangent that seemed related in his mind, such as how the digits in the decimal representation of pi were randomly distributed. RL readily volunteered all kinds of information about what he

thought and why, and his unprompted responses were lengthier in general than those of the other cases.

RL had taken Math 211 the prior quarter with Steve, and had taken a past college course in statistics at a different university. He also thought that both probability and statistics had been covered briefly in high school. Considering his own attitude at the start of Math 212, RL said: “As a future teacher, I look forward to mastering at least the basics.” Again reflecting his penchant for mathematical terminology, RL’s definition of what variation meant to him on the PreSurvey was “a measure of how a piece of data compares with the average of similar data.” His definition corresponds well to the idea of variation from the mean, and his example of something that varies was “sea level.”

There are two main themes for RL that I want to stress before sharing his thinking on the PostInterview: His consistent use of sophisticated mathematical and statistical language, and his dominant thinking regarding the expected value. Some examples of RL’s mathematically-flavored responses have already been shared, but I want to cite some more specifics because they really help define RL and also set him apart from the other students. The examples I’ll choose all come from the PreSurvey and thus represent RL’s style at the outset of Math 212. In the “Six Trials” of Q1c on the PreSurvey (using the Small Jar), RL explained his choices of “4, 5, 6, 6, 7, 8” by saying, “Reality does not obey the estimates of probability, so while 6 red candies remains the <u>average</u> outcome, variation is likely.” Later, in

reasoning about 50 trials at the Small Jar on PreSurvey Q3, RL claimed "a bell curve represents the most likely scenario – the extremes aren't seen often , the average is seen most often." As a final example, for "Six Trials" with the coin (Q7c on the PreSurvey), RL wrote that "while 25 flips are likely to be heads, in reality some variation is likely, so my numbers represent a range that averages 25." RL's responses above show his consistency in combining the usage of both average and variation in his reasoning, and he also persistently made sophisticated mathematical observations. For example, he expanded on his calculations with probabilities using decimals which were shared earlier:

> RL: Since I'm talking about decimals, when you multiply them they get smaller and smaller. But at least when you're using the 0.6 [for Red] it gets smaller more slowly.

Later, in the PostInterview, he actually referred to wanting to find the "inflection point" on a normal distribution of results. No other student in Math 212 used the kind of mathematical language with the same regularity as did RL.

While RL was unique in his mathematical fluency, RL was unique in another regard: He was the *only* student among 27 who took the PreSurvey and put an unqualified "Yes" when asked if the results of several trials would repeat (Q1b, "Several Trials"). One other student put "Yes" but then qualified her answer, but RL alone was mathematically blunt. Six reds were expected on one handful because "reds are likely to be chosen according to their relative percentage of the total," and six reds would come out every time because "returning the candies recreates the original conditions, so the odds

don't change." This latter response is very telling about the thinking of RL and the dominance of the expected value over his reasoning. He wrote similarly in comparing trials for the coin on the PreSurvey (Q7b "Compare Trials"), saying that "in the absence of any change of approach, the results [25 heads] are most likely to be the same." Finally, on the first PreInterview die question (Q9 "Sixty Tosses"), he emphasized his belief that "10, 10, 10, 10, 10, 10" would occur, repeating twice: "I think that's going to happen." RL's instinctual attraction to the expected value showed in his responses to both the PreSurvey and the PreInterview questions, as did his use of mathematical and statistical language.

Expecting Variation

Although the expected value was prominent in *what* he thought would occur, and proportional reasoning was included as a justification for *why*, RL showed appreciation for and awareness of variation in the common subset of PostInterview questions.

What was Expected: Concerning the expected value, when asked for the results of one trial at the Large Jar in Q1a on the PostInterview, RL said that multiple trials "would average a representative of 60 red." I again asked about just one trial, and he gave 60 red. Whereas on the PreSurvey RL had said that several trials would identically be the expected value of 6 reds, on the PreInterview he softened his expectation by saying "I think you're going to get it [6 reds] more often than not." On the PostInterview, with "Several Trials" at the Large Jar (Q1b), RL said that "I think I'd get more 60s than anything else."

However, for Q1c (“Six Trials”), he seemed to deliberately avoid choosing the expected value, instead putting “50, 55, 62, 65, 68, 70.” The theme that results should be close to expected value came through in RL’s responses to probability questions as well as in the sampling context. For example, in Q12 (“Compare Comments”), he found it “curious that there are between, I guess, 23 and 27, that there’s so few.” He also showed a sensitivity to his own prior arguments based on the theoretical ratio, offering a counter-argument to the idea that the spinner in Q12 was biased because no trials actually resulted in the expected value of 25 black:

> RL: Well, I don’t think that – just because nobody got a 25, that seems to me a little bit nit-picky, because you’re not – That’s adherence, that’s too close adherence to this principle of “It’s theoretical, and therefore that’s what I expect to see”

In other words, RL argued against the very kind of thinking he offered at times on the PreSurvey and PreInterview.

RL implicitly addressed the theme concerning repeated values by talking about the variation he expected, and by variation he generally meant values above and below the average. I noticed in the “Six Trials” at the Large Jar (Q1c), how his choices shared earlier were all different from one another, and so when he again had all different choices for “Six Trials” at the spinner (Q10c), I asked:

> DC: And he’s got 21, 22, 23, 27, 28, 29…Why those numbers?
> RL: Well, in either case, I did a pair, each equally far out from the mean in either direction. And none – there’s no repeats, but they’re all similar, but not identical.

I commented on how he had not listed any 25s, and he said: “Uh, no, because 25 is the theoretical expected result, but that’s not to say that that defines what happens…I don’t count on it.”

Besides themes concerning the expected value or how frequently results might repeat or be different, RL was clearly attuned to range expectations. In fact, at one point in discussing one trial at the Large Jar (Q1a), he even said: “Well, I am inclined to estimate a range, rather than give an exact number.” Later he did give a range for one trial, moving away from his expectation of 60 red to suggest results might be “within a pretty wide range...I would say, even as low as oh, 40 to 80, even.” On the spinner tasks of Q10, he went for a range expectation immediately, volunteering a range “somewhere between 21 and 29” for “One Trial” on Q10a. He also thought that a second trial (Q10b “Compare Trials”) would be similar, saying “I think that it’s likely to fall in a same range, similar range.” On the 35 Muffins question (Q8), RL mentioned the mean and mode but gave at least a small range for the expectation of the 36th muffin: “Well, if I had to guess, I would think it would probably be, uh, right in that center range – Probably 113 to 114.”

Why (Reasons for Expectation): Some of RL’s reasons for expectation used the language of possibilities and likelihoods that other cases had also used. For example, his response in Q10b (“Compare Trials”) showed he thought a similar range was likely, and on Q1b (“Several Trials”) he talked about how repeated results were likely to fall within a range tighter than 40 to

80 reds. When we were discussing Q2 (“Lists” for the Large Jar), he discussed the likelihood of getting all 60s in six trials:

> RL: Well, other things are also likely, even if they’re not MOST likely. And so, I don’t think that particular value – 60 in this case – is SO much more likely that you’re going to see it to the exclusion of all others, especially is such a small sample. If you did this 1000 times, you might see six 60s in a row…But the first six, to have no variation…{RL sounds doubtful}

RL also talked about the likelihood of not having a result of 25 black in twenty trials at the spinner (Q12 “Compare Comments”). For more than twenty trials at the spinner – I proposed he next think about 100 trials – if there had been no 25s then RL said he would “get a quizzical look on his face,” implying surprise or suspicion. On the PostInterview in particular, RL’s way of conveying a general sense of what was possible or likely was to talk in terms of outcomes being suspicious or not. In addition to “suspicious,” RL spoke of unlikely events as “iffy” and “fishy”. As another example, in considering Q2 (“Lists” for the Large Jar), RL regarded choice (i) as having too many high values. By saying “this is a lot of high values, and it just sets off my spidey-sense, I guess,” he was conveying the idea that extreme values were unlikely and their presence could be alarming. In the PreInterview, RL had used more of the same language as other cases in talking about possibilities and likelihoods, as when he discussed “One Trial” at the Small Jar (Q1a):

> RL: It is entirely possible that if there were 99 candies and 1 yellow, you could pick that one yellow every single time. It is possible. It’s on the far end of a bell curve, it’s extremely unlikely, but it COULD happen.

In the PreInterview, RL would often follow up his ideas of general possibilities and likelihoods with calculations, and he was quite adept at proportional reasoning. He was very explicit in describing why he predicted 6 reds for Q1a at the Small Jar:

> RL: For any given piece of candy, there's a 60% chance that it will be red, so when you multiply the decimal equivalent 0.6 per piece of candy times ten total pieces…I was trying to put a probability on any ONE given piece of candy, and then multiply that same probability out across ten

For the common subset of PostInterview questions, RL hardly referred explicitly to ratios or proportions hardly in his explanations, but his powers of reasoning proportionally had not diminished. That is, I did not think RL had lost the ability to calculate over the quarter, but instead of calculating aloud as he had done earlier, RL had more oblique references to the "theoretical prediction" on the PostInterview.

Rather than proportional reasoning or possibilities and likelihoods, the theme that was most strongly reflected in RL's reasoning for *why* he held his expectations had to do with variation, including the shape of the underlying distributions. For example, in "Six Trials" at the Large Jar (Q1c), he said "I've got a range between 50 and 70 is what I've got," explaining his choices by adding that "they are near the most likely value, but still wide enough to account for variation." RL referred the need for a symmetrical distribution for his six choices in the spinner task of Q10c, although he also implied a lack of surprise if the mode shifted away from the expected value of 25 blacks. On

two PostInterview questions not in the common subset, Q2 and Q11 ("Lists" for the Large Jar and for the spinner), RL frequently talked in terms of range, variation, and the shape of the distribution. For example, on Q2, here is how he talked about two lists that he didn't like:

> RL: Choice (v), that's an awfully narrow distribution [61, 66, 62, 62, 60, 59], and so I'm a little suspicious of that. And I'm suspicious of number (vi) for the opposite reason: It's all over the map [30, 10, 90, 20, 60, 50].
> DC: How do you mean?
> RL: The distribution is extremely wide, and so…it looks a little iffy to me

For a list he did like on Q2 ("61, 73, 56, 69, 59, 48"), he said: "With (ii), there's variation, but it doesn't seem extreme." He also liked choice (ii) "because the graph of this distribution skews to the left," as he thought it should. However, the scenario in Q11 prompted RL to "expect a symmetrical distribution." He favored choice (ii) – "26, 32, 22, 29, 24, 19" – because "there's a range that seems legitimately wide, and it also looks at first blush to be relatively symmetrical." Just as RL was consistent in reasoning about variation for the lists he did like in Q2 and Q11, he also had similar comments about a list he did not like in Q11, choice (v) – "24, 25, 26, 25, 24, 26" :

> RL: Yeah there's variation, but there's so LITTLE variation, that it discounts the possibility that even though a wider distribution of values isn't AS likely, it is still SOMEWHAT likely, and so… I'm a little suspicious of such a tight distribution.

In the latter response, RL used the language of possibilities and likelihoods as well as picking up on the theme of variation and the shape of the distribution.

RL's responses in evaluating and comparing, graphs focused on average, range, and distribution. Also, in making conclusions about graphs, he talked about the level of detail provided by different types of graphs, and he used the theme of reliability and consistency in context to help make his decisions.

Evaluating and Comparing Graphs: It was apparent that RL was attending to the averages in when dealing with data and graphs, as he explicitly stated so when considering Q8 ("35 Muffins"): "Well, I'm looking at the mean {Points to the summary statistics}, and I'm looking at the mode, in this case, which really stands out." He pointed to the histogram, reiterating "again, this 113 mode is very salient." He didn't mention the average hardly at all in the next question, (Q9, "Two Bakeries"), but he had focused on the average in some other questions. For instance, on a question not in the common subset (Q3, "Real: 30"), RL remarked that "this looks like, well, we've got a mode of 60 which is what I would expect to see, so it looks believable."

More than average, RL's responses in evaluating and comparing graphs had to do with ranges and the distribution. In thinking about where his own (36^{th}) muffin might be in relation to the 35 Muffins of Q8, RL noted that "it looks like most muffins weighed about 113, 114," and he later called these two weights the "center range." Also on Q8, he reasoned about where the "middle 50%" of the data was, drawing information off of the boxplot. In the

"Two Bakeries" scenario of Q9, he talked about in terms of subranges for both bakeries:

> RL: The interquartile range is narrower [for West End], you can pretty much count on most muffins are gonna be within a certain range, whereas over here on the East, on Tuesday you might go on in and get a 90 gram muffin, and then go in the next day and get a [muffin] that's 50 grams heavier!

The range for the East End bakery actually goes from 88 grams up to 142 grams, but it was clear that RL was reasoning by focusing on ranges and distributions. He also showed a willingness to consider the context, by thinking about what the experience might really be like for a customer to visit the bakery on a daily basis. He later personalizes the situation as he makes his conclusion, to be shared shortly. With other graphs, such as Q5 ("Small & Large"), RL repeatedly referred to the distribution, commenting on its shape and how far out it extended on either end. Often RL accompanied his remarks pointing to the effects of doing more trials on the shape of the distribution, and I'll share more about this when discussing the *interpreting* aspect.

Making Conclusions about Graphs: RL talked in terms of reliability and consistency when comparing the two bakeries of Q9, saying that a person might "be more inclined to seek consistency" in their muffin weights, and that it "looks like the West Bakery is a little more reliable." In contrast to the West End, at the East End there was less reliability, and RL's language took on a personal tone:

RL: The variation [for the East] is SO much that, if I'm looking for reliability, if I wanna know what I'm gonna – to expect, then I don't wanna mess around wondering if I'm gonna get a huge honkin' muffin, or I'm gonna get a little sub-standard muffin.

Thus, his conclusion was that there was less uncertainty in wondering what he would get at the West End bakery, and it also seemed that he had taken the context of buying muffins into account in making his decision.

He also commented on the level of detail and subsequent usefulness of the different graph types in Q9:

RL: Well, I think that with the boxplot you're sacrificing information about either end. So, we see on the boxplot that there was – The high weight muffin is, you know, over 140. But it's not until we look at the other one [the dotplot] that we see that several of them were that high.

His impression of the boxplot as giving restricted information in Q9 was an echo of his thoughts earlier in Q8, when commented how the mode was not represented in the boxplot. He knew the boxplot was telling him the range and interquartile range, but said "outside of that middle 50%, there's very little that I can glean from what's going on". He continued voicing his reservations about boxplots:

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

There are a few patterns in RL's overall reasoning that suggest he would have not found boxplots as useful as histograms or dotplots. For showing results

from sampling or probability experiments, he had mentioned wanting to look for differences among the results, something which can't be discerned from boxplots. Although he did talk in other tasks about ranges and subranges (which boxplots do show to a degree), he also emphasized shapes of distributions (which boxplots obscure). Distributions were a key to his other conclusions about graphs, such as in Q5 ("Small & Large"), where he was "more suspicious of Class A then of Class B," saying: "It's remarkable how similar the distributions are between them." He went on to explain that he felt the graph for Class A looked "a little too expected. It's almost too perfect."

Interpreting Variation

RL did not volunteer very much about *causes* of variation except on the question which specifically asked for speculation about causes (Q6 "Reasons: Muffins"). Also, he gave fewer responses that reflected *effects* of variation than he had in the PreInterview, and I'll highlight this difference later in the discussion. The dimension which came through the strongest in RL's response was *influencing expectation and variation.*

Causes of Variation: It was clear that RL could generate plausible causes of variation in both the "Reasons: Max" question (Q6 on the PreInterview) and "Reasons: Muffins" question (Q6 on the PostInterview). Also, on the PostSurvey (Data & Graphs), he listed reasonable causes for variation in rainfall patterns between the two cities. However, he did not wonder aloud about possible causes unless asked to do so on the PostInterview.

Effects of Variation: The indirect effects that I had attributed to other cases were completely absent from RL's responses. That is, he did not have difficulty in making a choice, he never expressed the view that "You Can Never Know," and he wasn't inclined to think that one can't have much confidence in choosing between real or made-up data. I wondered if RL's repeated references to outcomes being "suspicious" was an effect of variation, since the variation seemed linked to uncertainty in RL's mind for those situations. For example, he would view extreme values as unlikely but possible, and he wanted to see some variation in the results but not too much. It seemed that, because RL had shown an instinctual attraction for the expected value and results close to that value, an effect of variation for RL was to cause him to be suspicious of extreme results. He linked the likelihood of extreme results to the number of trials, and that leads into the last dimension of RL's thinking.

Influencing Expectation and Variation: RL said more than any other case about not only what more trials might do to the distribution of results, but also what less trials would do. In Q1a ("One Trial"), part of his response shared earlier included the thought that the average of multiple trials would be the expected value: "Over time, if I pulled 100 candies, put them back, pulled another hundred candies… I think I would average a representative of 60 red , 40 yellow." He had held the following thought even more strongly in the PreInterview: Even if individual results might vary, the average should match the theoretical ratio. However, by the time we got to Q12 ("Compare

Comments"), it seemed that RL was more comfortable with the idea that the average for the twenty trials was not the expected value of 25 blacks, because "ONLY 20 sets" were done.

On Q1b ("Several Trials"), he combined several ideas related to the number of trials performed:

> RL: Well, the more times I draw, the more normal the distribution, I think I'd get more 60s than anything else, but the more you draw, then the wider the distribution as well. The more you draw, the more chance there is of getting an outlier, or an extreme value. So, I would think that the more I draw, I'm more likely to get…Well, over time I think I'm more likely to get within a tighter range [than 40 to 80], actually

At first I thought RL was concluding that results of a greater number of trials would have a smaller overall range than would the results of a fewer number of trials. However, based on all his other responses, particularly those having to do with distribution, it seems more likely that what RL meant was that data for more trials would likely be more concentrated within a narrower subrange. In his response above, he mentions thinking that the mode of more trials would be the expected value of 60 red, and he suggests getting a normal distribution (although elsewhere he mentioned the expectation of a skewed distribution). He also thinks about the tails of the distribution extending through repeated trials.

The attention to the shape of the distribution was a key feature of RL's reasoning, and he knew that less trials gave a less consistent picture then doing more trials. In Q12 ("Compare Comments"), he expressed a small

amount of curiousity over why there weren't any results at 25 blacks, and so few near 25:

> RL: So, with 20 [Trials], it's a little tricky to say, but it is curious that we don't see ANY really, we only see a couple, here in the middle. So I don't think that's going to last very long, I think it's probably going to be balance by a preponderance, a run on 25s later

He also made it clear that fewer trials gave a less informative picture of the underlying distribution:

> RL: So, when you do five, it might be here, here, here, here here. {He is pointing at seemingly random places along the horizontal axis}. And so you're gonna see NOT a very , uh, you're not going to see a very descriptive graph. The more you do, the more it is going to start to appear as you expected.

I asked him to show me what he expected, and he drew a normal curve, centered at 25, saying: "Well, I think, eventually, you do enough, you're gonna get your symmetrical distribution."

RL was very consistent in talking about the shape of the distribution, and he was able not only to reason about the number of trials and the connection to the shape, but also the relative sizes of the samples and populations in Q5 ("Small & Large"):

> RL: I might have expected a more awkward , I guess, configuration or graph in Class A, with the small sample size. This [graph] points to the theoretical distribution pretty accurately, and with such a small sample size, I might not expect to see that.

RL was comfortable with Class B (which had larger samples coming from the Large Jar) having a shape that more closely matched what he expected a theoretical distribution would look like.

Thus, the picture that emerges for RL in terms of influencing expectation and variation is fairly reasonable: With fewer trials, he expects some variation on either side of the average, but an unclear shape for the underlying distribution. More trials pulls the extremes away from the center and yet also clusters the values closer to the average. He did comment on the expectation for a skewed distribution for the sampling tasks and a symmetric distribution for the probability tasks, and also had some sense that the two classes in Q5 ("Small & Large") should not look so similar because they were based on unequal sizes of populations and samples.

Discussion

RL came into Math 212 exhibiting stronger math skills in general than most of the other students, but by relying on those skills in the PreSurvey and PreInterview, he tended to overemphasize theoretical predictions as an instinctual response. His skills and instincts continued to be with him in the PostInterview, but he seemed to have responses that were more tempered, showing more sensitivity to variation than to the average. Overall, through the quarter his responses showed more shifts in the *expecting* aspect, and more stability in the *displaying* and *interpreting* aspects.

For *expecting* variation, recall that RL had been unequivocal in the PreSurvey and PreInterview that as long as the ratio didn't change, the expected value should appear over and over for sampling and probability contexts. At times he had seemed to be taking a philosophical or theoretical sort of mindset, but even when I had asked him pointedly about what might

actually happen in six trials at the Small Jar in the PreInterview, he indicated a lack of surprise if all six trials resulted in 6 reds. Also, I had shared earlier how he seemed initially emphatic that all 10s would result in sixty tosses of the die. Even on the PostSurvey for Data and Graphs, his graph of average daily rainfall for June was one of two graphs that used a straight line, showing a marked lack of expectation for variation. On the PostInterview, however, his responses reflected a clearly different trend. It still suspected that RL harbored the instinctual notion of the theoretical prediction as the best thing to say in the sampling and probability contexts, and in the data and graphs context we clearly was still influenced by measures of average. However, as far as telling me *what* he expected, on the PostInterview he gave range answers at almost every opportunity, and he talked about expecting answers to differ more often than he had on the PreInterview. Also, in describing *why* he held his expectations he seemed to have a shift in language: Whereas on the PreInterview he made heavy use of calculations to justify his theoretical predictions, on the PostInterview he talked frequently about more subjective notions of results looking “suspicious” or “fishy.” He also used the specific language of possibilities and likelihoods more in the PostInterview, and his talk of “suspicious” results seemed to be an extension of viewing results as unlikely. A significantly stable theme in RL’s reasoning *why* was his synthesis of average and variation in his responses. Some examples from the PreSurvey were given early in the case study, and in the PreInterview and PostInterview he continued to integrate average and variation in his

responses. The contrast between his ability to talk about variation early in the quarter while simultaneously appearing to ignore variation in some of his predictions is addressed at the end of this section, because I believe it is connected to RL's *interpretation* of variation, discussed further on.

RL's reasoning was fairly consistent in *displaying* variation, in the sense that when discussing graphs he used the same kinds of arguments throughout the surveys and interviews. That is, he attended to averages as well as spread, and commented on the distribution within the range as well. It did not seem that boxplots afforded RL much extra opportunity to comment on the range or subrange (such as the quartiles), because he already was facile at reading graphs and picking up on how the data was distributed. He did use the interquartile range in comparing the two bakeries on Q9 in the PostInterview, however. RL also about consistency and reliability in the PreInterview much as he had done on the PostInterview. For example, in Q8 ("MAX Wait-Times") on the PreInterview, he conflated range, variation, consistency, and reliability:

> RL: Looks like the Eastbound is more reliable. The range of Wait-Times is pretty limited. Over here [Westbound]…it's a mixed bag. There's a lot of variation, it's not a consistent pattern. I would say that the Eastbound is more predictable, and less variation.

He had similar remarks when considering the East End and West End bakeries. Some of his comments on the PreInterview clearly showed his previous exposure to statistics, as his language on Q13 ("Compare Graphs"):

RL: Group B is tighter, and definitely holds to a center more, and there's a wider range, more outliers in Group A…whereas in Group B you just didn't see that kind of variance.

The above response was *before* any formal instruction on probability or statistics in Math 212, and no other student showed that degree of fluency with statistical terminology in the PreInterview. It was not a surprise to see stability in RL's ability to reason in the *displaying* aspect.

There was a shift in RL's responses concerning *interpreting* variation. The shift had to do with his notion of the average of trials being the expected value in sampling and probability situations. In the PreInterview and some of the surveys, he mentioned that results should match the theoretical expectations, but even if they were different, then the average of the results would be the expected value. A good example came after he revised his expectations for the "Sixty Tosses" question on the PreInterview (Q9) away from all 10s: "So, yeah, if you're going to see a range, the average of that range will be 10, but not every response will be 10." He was similarly succinct in the PostSurvey for Sampling when he explained his choices for six trials at the Large Jar (Q2c), writing, "The mean of the above data is 60." In the PostInterview, although he continued to make careful choices whose mean was the expected value, he did not comment as stridently about how the average of repeated trials would match the theoretical predictions. It seemed to me that RL had much greater appreciation for how means, medians, and modes varied in the PostInterview.

Much stability was shown in RL's consideration of the *influence* of the number of trials on expectation and variation. In both interviews, he had a good sense for what performing more or less trials would do to the distribution. On the PreInterview, for instance, here is an example of what he said for Q4b ("Compare 30 & 300") and for Q11b ("Six Trials" with the die):

> RL: (Q4b) I expect to see a certain bell curve, given more trials.
> RL: (Q11b) You get a hundred people doing this, you're definitely going to see the extremes pop up more often

Earlier in the case study I had shared some of RL's similar expressions from the PostInterview. A bit of added detail came through in the PostInterview, in the way that RL was more explicit about data clustering near the mean as well possibly spreading out more to extremes and reflecting the underlying distribution with increased number of trials. One of his final comments in the PostInterview, on Q13 ("Compare Graphs"), reflected well the sense RL had conveyed throughout the quarter: "And so, it's easy to see how more sets will start to normalize that distribution and approach the theoretical prediction.

A last theme of RL's that I want to comment on has what I think is a key to the progression of his reasoning from Pre to PostInterview. The theme is that of reality versus probability, something I considered as a direct *effect* of variation. RL helped me get a sense of this theme in the PreInterview, when he verbally explained to me why he had put all 6s for "Six Trials" at the Small Jar earlier on the PreSurvey Q1c:

> RL: The first thing I did for part c [Q1c], where how many do you {think you'll get}, and I wrote "6" in every one.

DC Oh, yeah.
RL Being very strict as in probability-dictated reality, as distinct from described likelihoods. And so I went back and, instead of 6 every one, this one 6 and then a 5 and a 7, and a 3...
DC So you changed it {on the PreSurvey} ?
RL I did change it, when I went back and I thought, okay, reality is going to impinge on the strict likelihood by a given thing

The notion that repeatedly came through in the PreInterview was how probability (or theory) might suggest one result, but what happens in reality may be something else. Other cases mentioned a similar theme, and for RL he continued to mention the theme of probability versus reality, although his references were more implicit on the PostInterview. For example, he talked often in the PostInterview about the "theoretical distribution" or what would happen "theoretically," as if to make the implication that theory and reality were not always a perfect match. In the PreInterview, he was even more explicit, as the above responses from Q1c show. What I think was going on with RL is as follows: He was clearly knowledgeable about theoretical predictions, and on the PreSurvey and PreInterview he relied on that knowledge to drive his responses. He also had a sense that actual data would be different from what theory suggest, but I doubt that prior to Math 212 he had much experience with gathering data and seeing what really resulted. I think that he underwent some self-correction in the PreInterview as he checked his own responses, and his mentioning the theme of reality as different from probability was a part of his process of metacognition. His process comes through most clearly on Q9 ("Sixty Tosses") and Q10 ("Who Cheated") on the PreInterview, and recall that on Q9, RL had not only put all

tens but emphasized that “I think that’s going to happen.” After seeing the student Lee’s supposed results of all tens on Q10, here is how RL reasoned:

> RL: It would be pretty funny for, in a world of imperfect scientific conditions, to see the likelihood matched so closely. {Laughs} This was like, when I originally did this [Q1c “Six Trials”] and I put 6 down for the first one, and then I’m saying “You know what? We’re living in the real world, this is not going to be 10, 10, 10…We’re going to be, here’s a 12, but here’s an 8…{He changes his six 10s to 12, 15, 9, 11, 8, 5}

Later, he explained his change of mind and does a very thorough job in his explanation, again blending many elements of statistical reasoning:

> RL: And the reason actually, the reason why I would not go for this [All 10s] after all is because you’re going to see a range of results. I’m changing my mind because I’m making the same mistake that I was accusing some kids earlier of, and that I was considering average but not considering variation. You need to consider variation to get the full picture. So, I think I was being limited in my consideration.

The reason I have chosen to share RL’s thoughts so extensively in this regard is because they really show his reflective process. It is true that when he first put all tens on Q9, I let him defend his choice and then I double-checked to make sure he understood the intent of the question was for him to put what he really thought might happen if we did the experiment in Steve’s class. RL then reiterated his reasons for wanting all 10s. His subsequent thoughts came prompted by his own reflections on the next question of “Who Cheated,” which shows me the power of the sequencing of good questions to prompt thinking about variation.

In summary, RL had a fairly sophisticated command of notions of probability and statistics and mathematics in general at the outset of Math 212. In terms of a sort of hierarchy for appreciating variation, RL quickly went from expecting no variation to the more basic notion that results would in fact vary. He was slower to appreciate the higher notion that averages can also vary. I think that RL was greatly influenced by doing the activities in Math 212, although he never once referred in the PostInterview to the simulations we did in class. What I think happened was that RL's instincts and strength at the start of the quarter were mainly theoretical, and he had an appreciation of variation in a cognitive way only. The class experiences reinforced what he began telling himself in the PreInterview – That reality does offer variation from theory. In the PostInterview, he expressed the theme of reality versus probability more implicitly than explicitly, because it was like a self-lesson that he had already learned and then seen confirmed in class.

Conclusion

The revised framework from the end of chapter five was useful as way of organizing and analyzing the six cases' responses to PreInterview and PostInterview questions. In letting responses from all six cases contribute to all the questions on the PreInterview, I was able to get a sense of how the students' thinking fit into the framework while at the same time gaining a deeper understanding of themes that seemed to be emerging within the framework. In other words, the framework helped me look at the students

responses, and what I saw in those responses helped me in turn better understand the framework.

Then, in looking at the PostInterview results, I chose a smaller subset of questions with which to focus on each individual case's conceptions of variation. Three of the six questions chosen from the PostInterview (Q1, Q8, and Q10) were isomorphic to questions asked on either the PreInterview or the PreSurvey, and the other three questions were unique to the PostInterview. Again I used the framework to organize and analyze the results, this time for the purpose of getting a detailed picture of individual conceptions and looking for stability of shifts in thinking from over the quarter. In the next chapter, I'll summarize the research findings in terms of the framework for understanding, and discuss comparisons of the six cases to each other.