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

Individual Cases

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

Of the eleven subjects interviewed, there were three females (SP, EM, and DS) as well as three males (GP, JM, and RL) who were selected to be the six case studies

for this research. They were selected mainly because their collective responses spanned all the themes of the framework. Moreover, they had no problem sharing their thoughts in the interviews, and their narrative provided vivid illustrations of their thinking. Each of the six cases participated in two videotaped interviews, with each interview lasting about 45 minutes. The PreInterview was given within two weeks of administering the PreSurvey, and was conducted before formal instruction on probability and statistics in MET 2 had begun (recall that the first four weeks of the quarter in MET 2 was spent on geometry). After four weeks of doing lessons and activities on probability and statistics (this time frame corresponded with weeks 5 – 8 in the ten week quarter), the PostInterviews were conducted. To illustrate the comparisons for each case, I'll mainly use responses to a subset of questions from the Pre and PostInterviews.

The interview questions were summarized in Tables 4 and 8 of Chapter Three, and they are found in their entirety in Appendix B. I chose a subset of the interview questions for case analyses for three reasons. First, the cases' collective responses on these questions spanned all the themes of the framework. Second, the questions themselves spanned the three contexts of sampling, data and graphs, and probability situations. Third, the questions were specifically constructed to so that PreInterview questions were isomorphic to PostInterview questions. By isomorphic, I mean that the questions were phrased similarly or addressed similar ideas. I've reorganized Tables 4 and 8 to show how questions matched up, and also to assign nicknames that will help identify the questions used in this section (see Table 12).

Table 12. *Isomorphism of Interview Questions*

Pre	NickName	Scenario Involved	Post	NickName	Scenario Involved
Q1a Q1b Q1c	One Sample Several Samples Six Samples	Small Jar : 60R/40Y (Samples of 10)	Q1a Q1b Q1c	One Sample Several Samples Six Samples	Large Jar : 600R/400Y (Samples of 100)
Q2	Compare Lists		Q2	Compare Lists	
Q3	Graph: 30		Q3	Graph: 30	
Q4	Graph: 300		Q4	Graph: 300	
Q6	Causes: Train	21 Times Recorded (One MAX Train)	Q6	Causes: Muffin	20 Weights Recorded (One Bakery)
Q7	Compare Graphs		Q7	Compare Graphs	
Q8	MAX Wait-Times	10 Wait-Times Each (Two MAX Trains)	Q9	Muffin Weights	12 Muffins Each (Two Bakeries)
Q9 Q10 Q11	One Sample Who Cheated? Six Samples	Six-Sided Die (Samples of 60)	Q10a Q10b Q10c	One Sample Compare Samples Six Samples	1:1 Spinner (Samples of 50)
*	*	*	Q11	Compare Lists	
Q13	Likelier Graph?	2:1 Spinner (Samples of 60)	Q13	Likelier Graph?	

(* Along this row, the Post Q11 "Compare Lists" was in the Probability context and did not have a direct counterpart in the PreInterview. Post Q11 is similar in structure to Post Q2 and Pre Q2.)

In Table 12, I've created nicknames that reflect the content of the questions, and in general the Pre and Post questions can be matched by their nicknames. For example, "One Sample" for the Small Jar on Pre Q1 is similar to "One Sample" for the Large Jar on Post Q1. "MAX Wait-Times" (Pre Q8) is similar to "Muffin Weights" (Post Q9), and "Who Cheated?" (Pre Q10) gets at the same essential idea as "Compare Samples" (Post Q10b). There is one question in Table 12 (Post Q11) which does not match directly across to a counterpart in the PreInterview. Post Q11 had subjects "Compare Lists" in a probability context, but on the PreInterview I only had subjects "Compare Lists" on Q2 in a sampling context. Despite differences in context, useful comparisons of subjects' responses were still made between Pre Q2 and Post Q11.

In next presenting the case studies, I'll use the following structure. First I'll introduce the case, summarizing upfront the main points of stability or shifts in a student's thinking. Then I'll describe further details according to each aspect of the framework.

The Case of DS

DS was a very energetic individual who readily expressed opinions and thoughts on all the questions. She had taken MET 2 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.

Summary: It was clear from the PreSurvey and PreInterview that prior to the class interventions, DS already had a good grasp of the basic ideas involved in probability and statistics. She showed a facile use of proportional reasoning, and usually expected results of repeated samples to vary. She also gave reasonable ranges for predicting results of six samples, but was generally wide on her ranges for thirty or more samples. She expected ranges to increase as the numbers of samples increased. Lastly, in attending to graphs DS referred to center, range, and shape of the distribution.

DS corrected herself at two key points during the PreInterview: Once when she first thought that all tens was a good guess for "One Sample" of the die tossing (Pre Q9), and again when she initially thought Group B had a realistic graph in

"Likelier Graph?" (Pre Q13). In both instances she changed her mind on the basis of her "Won't be Perfect" reasoning strategy. I was surprised that she misidentified "Graph:30" as actual results for Pre Q3, since I would have expected her to say that it also looked too "perfect."

The main changes I noticed in DS's collective responses were that she had more complex responses in the PostInterview. In particular, she tended to use more descriptive language in the PostInterview than in the Pre when talking about the variation in situations, such as how data was "clustered". She attended to more features of the distribution (average, range, shape, and spread) in the Post than in the Pre, often incorporating several features in a single response. DS had some fairly reasonable ideas and good ways of communicating during the PreInterview, but she added depth to her responses and expressed herself even better in all aspects during the PostInterview.

Expecting: I'll use DS's responses to PreInterview Q1 as a starting point for this discussion. In "One Sample" from the Small Jar, DS predicted a result of 6 reds:

I: [Pre Q1a] How many red do you think you're going to get?
DS: I think I'm going to get 6 red.
I: Why do you think that?
DS: Because 60% of the 100 are red, and 6 is 60% of 10. So, it's not for sure. The odds are.
I: What do you mean, "it's not for sure" ?
DS: Well, 'cause, I could get a different amount of reds, and a different amount of yellows.
I: Right now you're saying 6
DS: Six is the best shot.

Her response was very reasonable, and she reasoned proportionally yet also acknowledged the possibility of variation. In "One Sample" from the Large Jar, she

predicted a value close to the expected value of 60 reds:

I: [Post Q1a] How many do you think are red?
DS: Sixty-four
I: Alright. Why do you think that?
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.

She again used proportional reasoning and considered the possibility of variation, and she used the same "odds" language as in the PreInterview. She actually expected variation from the mean, and had a prediction which she knew was "close" enough to be reasonable. What I found interesting was that she explicitly said that she probably wouldn't get 60 reds, and she gave a reason. While the expected value of 60 reds may be the most likely outcome for one sample, DS's response seemed to acknowledge that the likelihood of actually getting 60 reds is small.

For predicting "Six Samples" from the Small Jar on Pre Q1c, DS had a reasonable list of "4, 5, 6, 6, 7, 7", and I wondered if she would just multiply by ten when moving to the Large Jar. However, on Post Q1c her "Six Sample" predictions were "56, 58, 60, 61, 62, 64" , which all fall within a reasonable range for sampling from the Large Jar. In the probability context, however, DS initially expected no variation when considering the number of times each face of the die would show in "One Sample" of 60 tosses (Pre Q9):

I: [Pre Q9] What do you think is going to happen, for these faces?
DS: I'll just go, ten of each [Writes down all tens]
I: Why do you think those numbers are reasonable?
DS: Because… one out of 6 is going to roll up a "1", and one out of 6 will roll up a "2"…But then, going back to that question about picking the colored candies and the 6,6,6,6… That's kind of … 10,10,10,10 is kind of like 6,6,6,6… so it probably will vary somewhere.

I: Well, put what you think, Debbie.
DS: Ten is as good a guess as any.
I: If you rolled it 60 times, that's what you think you're going to get?
DS: Sure. [Seems pretty confident] It's as good as any guess.

DS clearly was influenced by proportional thinking, but I noticed that DS reflected back to "Six Samples" from the Small Jar. Her expectations from Pre Q1c seemed to conflict with her expectations for Pre Q9. However, an outcome of ten was "as good as any guess" for one face, hence good enough for all faces.

When I showed DS the supposed results of dice tossing on Pre Q10 ("Who Cheated?"), she was quick to identify Lee's list of "10, 10, 10, 10, 10, 10" as unbelievable:

I: Explain your reasoning, please.
DS: Well, because really, the 10, 10, 10, would be so unusual that it would come out that way.
I: Ok, it would be so unusual, and yet that's exactly what you said you thought might happen [Turns back to Q9]
DS: Well, I don't really think that it's going to happen. It's a guesstimate… It's an educated guess.

For DS, all tens was a reasonable guess and she listed all tens as an a priori expectation. When faced with the same result of all tens as a supposedly a posteriori result, she was quick to see that all tens was just not very realistic. She went back to Pre Q9 to change her "One Sample" result list from "10, 10, 10, 10, 10, 10" to "6, 8, 10, 10, 12, 14", saying all tens was too "perfect", and unlikely to happen in real life. I'll comment more on her line of reasoning against reality being "perfect" in the *interpreting* aspect. When asked to evaluate Lynn's list ("10, 11, 10, 10, 9, 10") as a part of Q10, DS said:

DS: And then Lynn, I don't think that out of 60 rolls that there's not enough variation, between what came up how many times...he [Lynn] only went over one and under one. Where really, chance could probably have a broader range.

I thought DS had some good reasoning in evaluating the different lists in the, and she clearly connected a narrower range with having less variation. When I asked DS to predict how many fives would result in each of "Six Samples" of sixty tosses (Pre Q11), she wrote "6, 8, 10, 12, 13, 14" saying she "liked those numbers". She also indicated she thought that with more samples, she would have chosen a broader range. The tendency of wider ranges in larger samples also arose for DS in the *interpreting* aspect.

DS never again listed a string of all identical numbers when given the opportunity to predict results. For instance, in the probability context on the PostInterview, she listed a reasonable "21, 23, 25, 26, 28, 29" for "Six Samples" of the spinner (Post Q10c). When explaining her choices, she included some proportional reasoning, and then said:

DS: So I have one 25 here. And then I have a few scattered close to 25, but not 25...'Cause there's gonna be variation, because the spinner CAN land anywhere, but probably on average it'll be close to 25.

DS included many themes in her above response. She appealed to the notion of distribution by describing how results are "scattered close to" the expected value, and she acknowledged that individual results and the average of a set of results will vary.

Displaying: DS suggested that the thirty supposed results of "Graph: 30" on Pre Q3 were actual results:

I: [Pre Q3] Which of the following do you think is most likely?
DS: Oh, I think… those could have been the [actual] results
I: Why do you think that's the most likely?
DS: Because six is our odds-on favorite, and they just didn't have a lot of variation when they picked out their candies.
I: What makes you say "they didn't have very much variation" ?
DS: Because here's six, and they're only one away from six, on each side [She pointing with her finger on the graph, tracing out the range]

Thus, DS was comfortable with the unlikely narrow range portrayed by "Graph: 30" (see Figure 22), but she didn't have too much to say in terms of her justification. She focused briefly on the mode of 6.

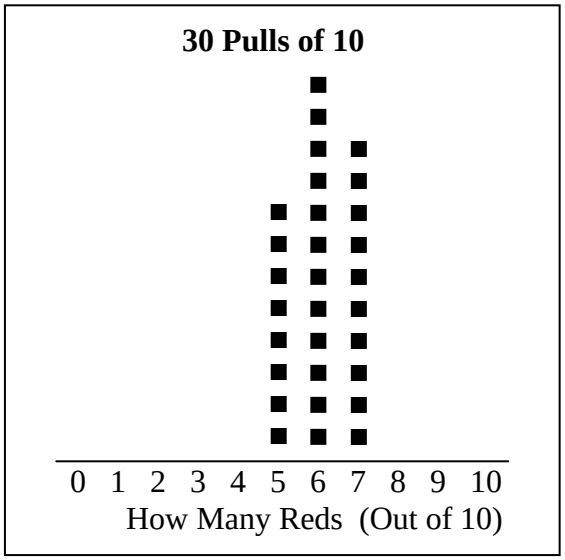

Figure 22 – PreInterview Q3 "Graph: 30"

She also attended to the range in making her evaluation, noting how the thirty results ranged from 5 red to 7 red candies. It is possible for thirty actual results from the Small Jar to look like the graph shown in "Graph: 30", but not very likely.

In the PostInterview, the "Graph: 30" for the Large Jar on Q3 really did represent actual data (see Figure 23), and DS made a correct identification of the graph as being authentic.

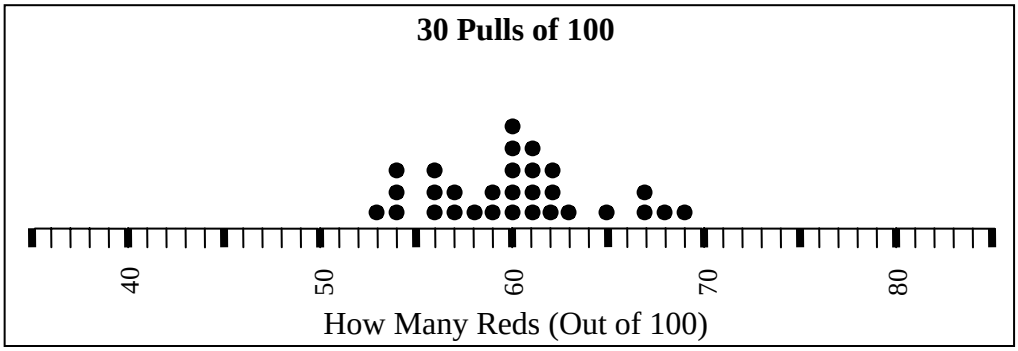

Figure 23 – PostInterview Q3 "Graph: 30"

More importantly, she offered better reasoning on the Post than she did on the Pre:

I: [Post Q3] So what do you think is most likely... that Craig's class just made up those results, or those are the actual results…
DS: No, I think those could be the actual results. [She is quick to respond]
I: Could you explain why you think [that's] most likely?
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…
I: So you like that
DS: Yeah, and then there's just, you know, a little straggle here and a straggle there [She marks the min and max]

DS's evaluation of the graph included a focus on the mode of 60. When she said that 60 is the "average number of reds" she means that 60 is the expected value, not the mean of the data set shown in "Graph: 30". She also appealed to the spread of the data by talking about the "cluster" of results around 60. Finally, she attended to the extreme values by marking them on the graph. Thus, her multi-thematic response on Post Q3 involved three themes focusing on average, range, and spread.

When DS worked on the "Compared Graphs" task in Pre Q7 (see Figure 24), she suggested that both graphs told a similar story about the duration of the train trip, "that it was somewhere between 58:30 and 59:30".

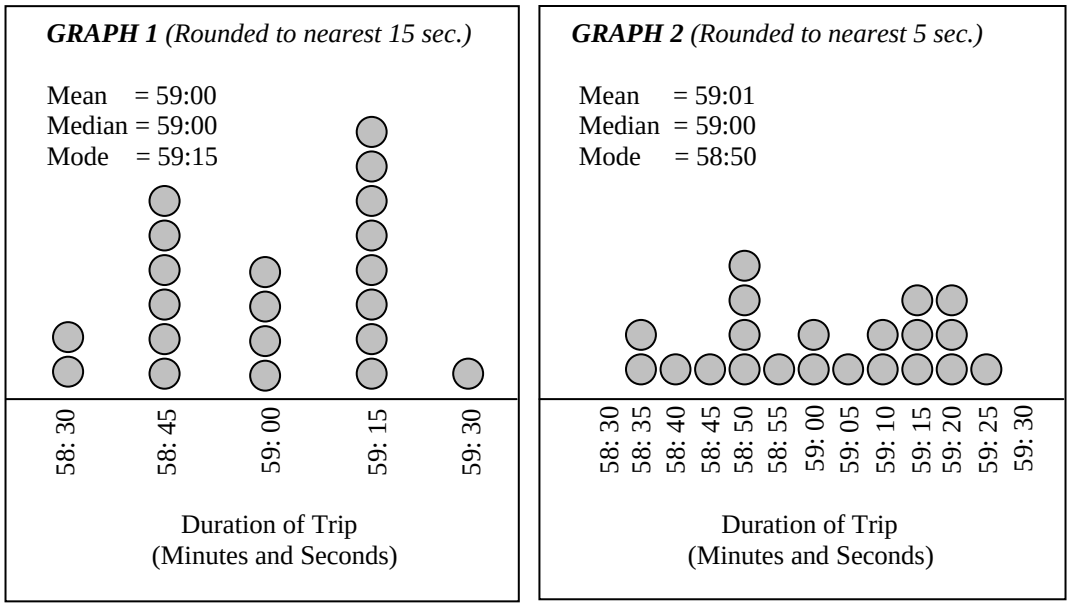

Figure 24 – PreInterview Q7 "Compare Graphs"

Thus, her initial focus was on the range. For Graph 1, she pointed to the subrange represented by the two middle bars and said that "more people experienced that time frame", and she also counted individual data points on both graphs to help her compare. What I found most interesting about DS's response in Pre Q7 was her conclusion about the two graphs that "I think I like them both. Either graph is fine." Her response on the similar "Compare Graphs" on Post Q7 involved more distributional reasoning and also a firmer decision in favor of Graph 1 (see Figure 25). She thought the two graphs on the Post Q7 told her different stories, with Graph 2 appearing more spread out:

> DS: Yeah, 'cause this [Graph 2] is kind of...it's like spread out in teeny increments, and kind of detailed like that. And also, in this Graph [2] there's so many numbers that you kinda go "Too Much!"...Like too much flatness, for it to really make a statement about how much it weighs. Where here [Graph 1] it makes more of a statement, like "Oh, probably weighs 109 or 111 – Somewhere in there."

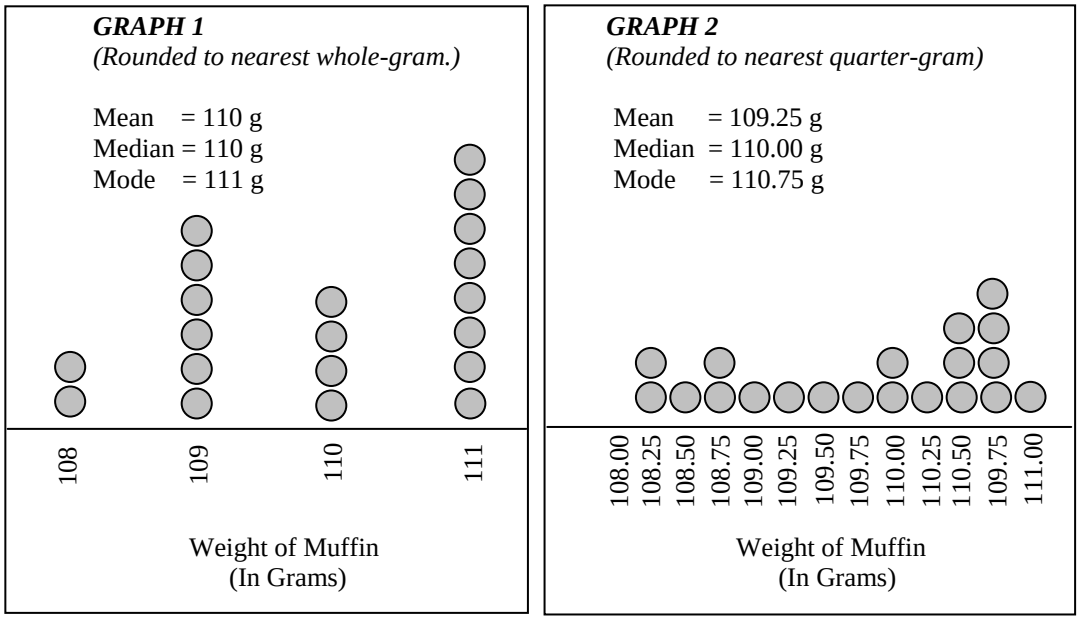

Figure 25 – PostInterview Q7 "Compare Graphs"

DS has two dimensions reflected in her response. For evaluating and comparing graphs, she attended to the shape of Graph 2, noting its "flatness". For making conclusions about graphs, she emphasized the level of detail and subsequent usefulness of Graph 1 versus Graph 2. Specifically, even though the rounding is finer in Graph 2, DS liked Graph 1 because it better conveys to her where most of the data fell.

Another example for DS's reasoning about *displays* of variation comes from the "Likelier Graph?" questions on the Pre and Post. On Pre Q13, Group B's graph is fake (see Figure 26), but DS initially said "I think Group B looks more like what I would expect." When asked why, she appealed to the shape, saying "it's that famous curve" (the graph was roughly bell-shaped).

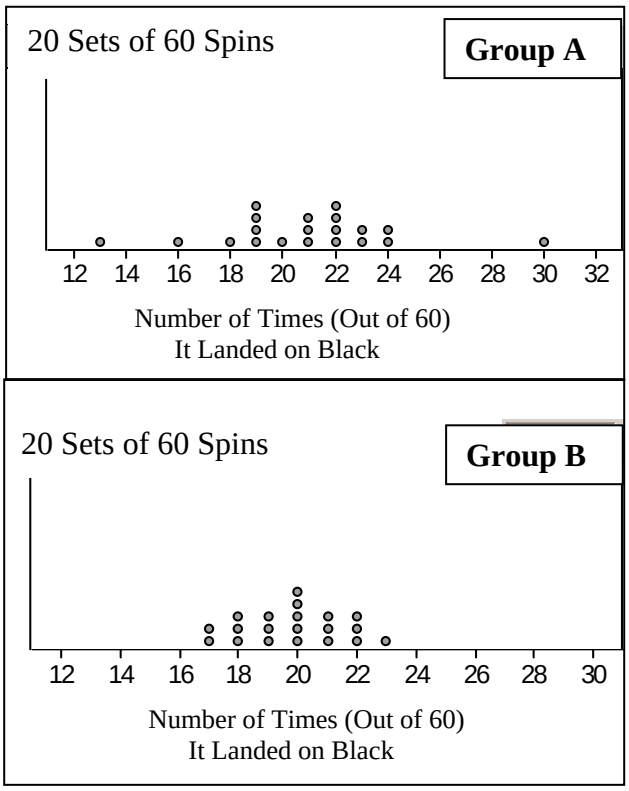

Figure 26 – PreInterview Q13 "Likelier Graph?"

She also focused on the mode, which was at the expected value of 20 blacks for a sample of 60 spins of the 1:2 (Black:White) spinner. Her comment was:

> DS: And so, most of those, in Group B, fell in that one-out-of-three...20 [blacks]. And then, you just vary a little on each side, 'cause you only have 20 sets [of 60 spins per "set"]. You don't have a lot of sets.

Finally, she focused on the range for Group B, which she liked. DS used all elements of the distribution – average, range, spread, and shape – in her evaluation of Group B, and she was convincing herself that the graph was reasonable. Then she changed her mind:

> DS: Back up. I think Group A looks more real.
> I: Now what are you thinking?

DS: Well, now I'm thinking that, you know, that it's not going to always end up in this perfect graph picture. So this [Pointing to Group B] would be, if you were going to fake a graph? This would be a fake graph [Laughs].

When pressed for more reasoning about her change of mind, her main rationale was the expectation for a more expanded range with 20 samples than what was pictured for Group B.

I thought that DS had some reasonable thoughts on the "Likelier Graph?" task in Pre Q13, and she expressed her ideas well, but in the PostInterview she expressed herself even better on this task (see Figure 27).

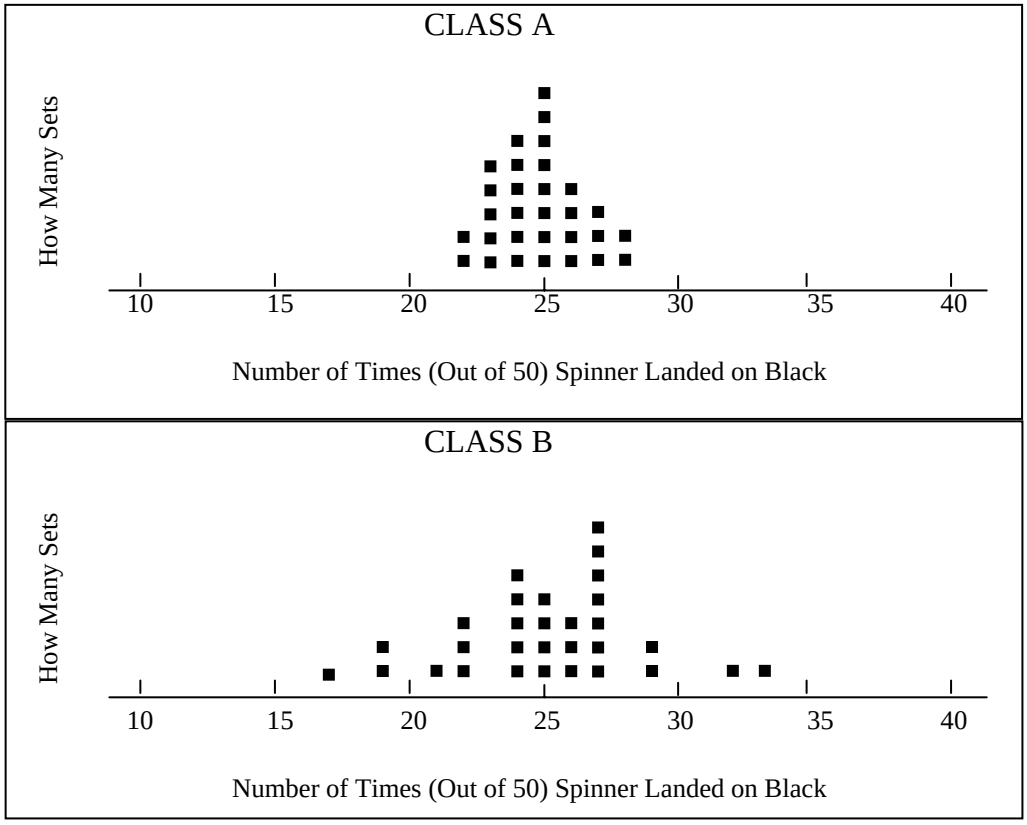

Figure 27 – PostInterview Q13 "Likelier Graph?"

This time Class A had the fake, more compact bell- shaped graph, and Class B was real:

DS: There's this...kind of, you know, bell curve [for Class A] that's kind of clustered right in the center, without much variation on either side of that 25 [The expected value]. Class B, you still have your cluster in the center...and then you have a few little odd birds [The extreme values]

DS considered shape and spread as she compared the two graphs (see Figure 27). Class A's mode was right at the expected value of 25, whereas Class B's mode was at about 27, but what DS particularly liked in Class B was the "cluster in the center". She also liked how Class B had a broader range than Class A's range of about 22 to 28. She was quick and confident to denounce Class A as a fake:

I: Do you have a sense that one class or the other is likelier to have...
DS: [Interrupts] I think Class A cheated.
I: Well, why do you think that?
DS: Well, because it's toooooo perfect. I just think somebody would have gotten, you know, under 20, or 20, or you know, 32...You know, even if it's just one person, that would be out of the cluster.

She shows her a sense of distribution by wanting a grouping close to the expected value, tapering off on either side of that value, and a reasonable sense of range.

Interpreting: DS continually referred to the "perfect" results in her reasoning on both the PreInterview and PostInterview. I saw that her sense of the "perfect" world was tied to her perception of variation. As seen in her above response to Post Q13, when rejecting Class A as the "Likelier Graph?", she claims Class A looks "toooooo perfect". Class A's graph is not completely symmetrical, but what DS meant was that the general shape and tightness of range was not realistic. The same idea of the "perfect" shape comes through in her comment about Group B's graph in Pre Q13's "Likelier Graph?", when she says that one won't always have a "perfect graph picture." Group B did have the mode at the expected value of 20 blacks, and

DS was explicit about how "your odds are...1 out of 3 is going to be black, in a perfect world."

DS's reference to the "perfect world" shows how she connects perfection to the absence of variation. She seemed to perceive probabilistic theory as predicting what would happen in the "perfect world" , while results in the real world varied away from the "perfect". I asked her to explain more about her perceptions when she had decided that rolling a die sixty times and getting each 10 of each face was unrealistic:

> DS: [Pre Q10: "Who Cheated?"] Because it's TOO perfect [Lee's choice of all 10s]. Life doesn't happen that way. It could but it doesn't. [Laughs]
> I: Why doesn't it?
> DS: Because it's… a random thing.
> I: Could you tell me what you mean by that?
> DS: That there's chance involved, so, whenever there's chance, then things won't necessarily turn out perfectly. Like, in a perfect world situation, where the dice was loaded.

I think DS's reasoning serves her well. If it were a "perfect world situation," she seems to be saying, then less variation would mean more consistently correct predictions. Chance leads to variation, both of which are related to uncertainty. The "perfect" result was clearly one number, she explains further:

> DS: It is an idea that I hold. So, 'cause I think that, um, that there IS the chance that it'll come up perfect, but there's … "perfect" is one [Holds up hand to signify one number] , and there're more things that are imperfect, like, not perfect. [Waves hands to show distribution of other numbers on either side of the "perfect" number] So, there's a lot more options for the imperfect.

Her explanation above shows why DS said for "One Sample" at the Large Jar in Post Q1a that her result probably wouldn't be exactly 60 reds. A sample result 60 reds is just one of many other possibilities. Her sense of the "perfect" went beyond just the expected value to include shape and spread, as exemplified by her responses to the

"Likelier Graph?" questions. She also commented on a list of supposed results for six samples of the fair spinner in "Compare Lists", Post Q11. About list (v), she said:

> DS: Choice (v) is good, the only thing is that it's so…Perfect...You know, 24, 24, 25, 25, 26, 26…There's not a lot of variation there, which I think there might be a little more.

List (v) is so "perfect" because it only varies by one on either side of the expected value of 25, and because it shows a uniform distribution of two samples for each of the outcomes of 24, 25, and 26.

Another example of DS's thinking in the *interpreting* aspect is how she built upon her notion of what effect taking increasing numbers of samples would have. She already thought in the PreInterview that more samples meant a widening range and she held onto that notion in the PostInterview:

> DS: [Pre Q1b: "Several Samples" of the Small Jar] The more I choose candies, the more chance there will be that I'll get different than six reds. Either fewer or more.
> DS: [Pre Q11: "Six Samples" of the Die] If we had more sets of 60, then I would make my numbers go lower than…[Showing with hands a greater range]
> DS: [Post Q13: "Likelier Graph?"] The more spins you do, I think there's more chance that you'll get… A number that varies from your, you know, further from your 25.

DS had already expressed how more samples meant more chances for the "imperfect", what happened in the PostInterview was that she added to her notion of more samples meaning a wider range. In the Post (but not in the Pre), she added the idea that more samples gave more chances to actually attain the expected value. For example, on Post Q10c, "Six Samples" of the spinner, DS said: "But the more times you spin it, the more chance that you'll get 25." Now compare what DS said above in "Several

Samples" of the Small Jar on Pre Q1b to what she said about "Several Samples" of the Large Jar on Post Q1b:

> DS: [Post Q1b] 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 number.

Whereas in Pre Q1b she mention getting "fewer or more" with more samples, in Post Q1b she includes the language of "variations" to describe a widening of the range, and she also added the distributional idea that more samples meant more near the center, "around" the expected value of 60 reds.

Finally, DS also related more samples to the shape of a distribution in the Post, but not in the Pre. A good example comes from "Graph:300" (see Figure 28), and for the Small Jar on Pre Q4 she mainly attended to the range:

> DS: [Pre Q4] Well, this ["Graph:300"] has a broader range of picks, of number picks [Her finger traces out the range on the horizontal axis]

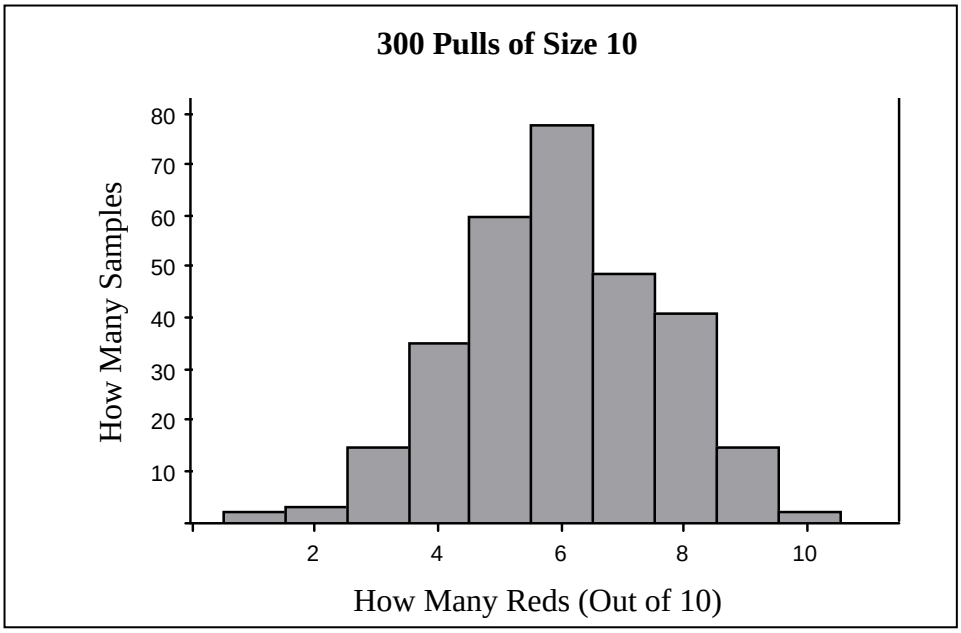

Figure 28 – PreInterview Q4 "Graph: 300"

Her answer was much more thorough on Post Q4 in evaluating the "Graph:300" (see Figure 29):

> DS: [Post Q4] Well, the most common is right around 60, and then there's fewer on the edges as get further away from 60. And, in class when we did, on the computer, the more pulls you do, the more evenly shaped your graph is going to be. Where fewer pulls, you're going to have a little more unevenness in your curve

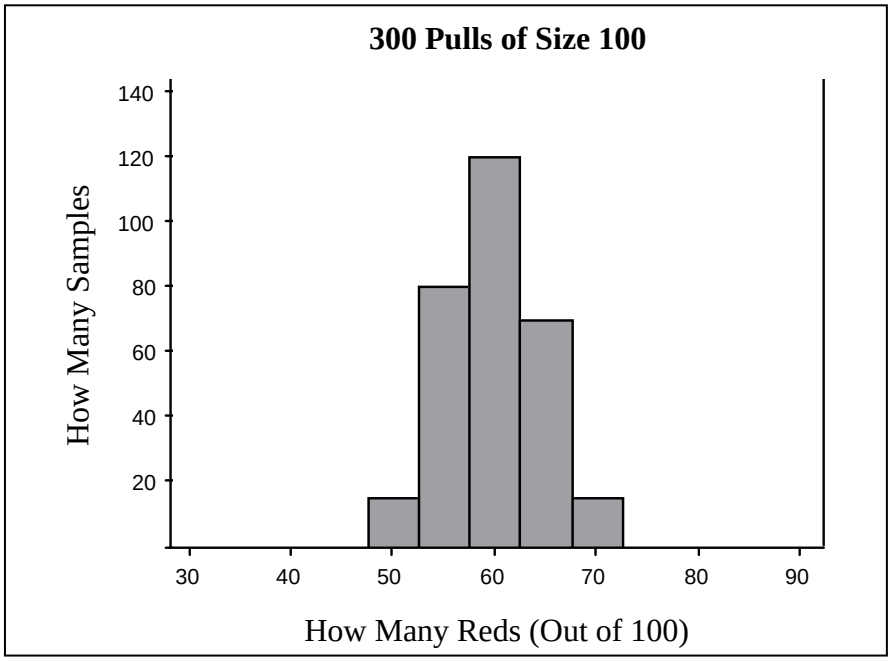

Figure 29 – PostInterview Q4 "Graph: 300"

DS's thoughts are extremely well-said, and she includes many themes such as attention to average, range, and she includes class experience in her reasoning. The focus for the *interpreting* aspects is the effect of the number of samples: More pulls means a more "even" shape, less pulls means more "unevennesss" in shape.

In conclusion, there was one instance on the PreInterview where DS had predicted all tens for each side of the die as the outcome of 60 tosses on Pre Q9 (One Sample). During the PreInterview, she changed her mind about the reasonableness of

her prediction, and never again predicted identical results. She erroneously supported fake graphs as being realistic on the PreInterview, but her later evaluations of such real-versus-fake graphs were more reasonable on the PostInterview. DS had stability in her language and reasoning from Pre to Post about how results wouldn't always be "perfect", and added to her thinking about the effect of doing increasing numbers of samples in the PostInterview. The biggest difference for DS in all aspects from before to after the interventions was a qualitative broadening in her reasoning skills. She had some reasonable ideas in the PreInterview, but her responses were deeper, more complex, and better expressed by the time of the PostInterview.

The Case of GP

GP was an effusive character who used gesture quite frequently 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."

Summary: In *expecting* variation, on the PreSurvey and PreInterview GP showed a fairly naive understanding of how unlikely some extreme values were. He became more sensitive to the presence of outliers by the time of the PostInterview, and incorporated more language about what was possible or likely. He consistently thought results from multiple samples should usually be different from one another,

and also mentioned experience as a reason on both Pre and PostInterviews. The sampling activities seemed to make an impression on GP, particularly the computer simulation.

When considering *displays* of variation, on the PreInterview he made many reasonable evaluations and comparisons, and he continued to do so in the PostInterview. He was quicker to make conclusions about graphs in the PreInterview, and was occasionally less decisive initially in the Post. The biggest change for GP in this aspect was that he became much more sophisticated in his communication about graphs, and on the PostInterview he reasoned more distributionally.

For *interpretations* of variation, GP consistently focused on the physical nature of the candy mixing, although he did so to a lesser extent in the PostInterview. He also consistently used forms of the word "random" to describe many facets of variation. While he knew that the number of samples was likely to extend the range of results, this idea came out stronger in the PostInterview than in the Pre. He also considered the numbers of candies in the jar to be influential for the PostInterview.

Expecting: On the PreInterview, GP gave "6 red" as his prediction for "One Sample" of the Small Jar (Q1a), and he gave a proportional reason. Thus, I knew that GP was capable of reasoning proportionally, which was a significant finding because he also frequently relied on additive reasoning (meaning that he focused on the sheer numbers of candies). He included additive reasoning (to be discussed further in the *interpreting* aspect) when he predicted "70 reds" for "One Sample" of the Large Jar on Post Q1a. His language in answering "Several Samples" was similar on both Pre and

Post Q1b, when I asked him if he would get the same results each time:

> GP: [Pre Q1b] Probably not. Probably not, no.
> GP: [Post Q1b] No. I mean, not every time. It could happen, but...Not very likely.

Both of his responses show the theme of possibilities and likelihoods, and on the Post he often was very explicit about results being possible but unlikely. For "Six Samples", GP's range was too wide on the PreInterview (he picked 1, 3, 4, 5, 6, 10), and he said:

> GP: [Pre Q1c] It just randomly came to me. I thought if you stick your hands in there randomly, you could just pick up any number, between 1 and 10

The result of 1 red is very unlikely for "Six Samples" of the Small Jar, yet GP did not comment on the relative likeliness for that lower extreme value. He also did not seem to consider zero reds as a possibility, and he uses "randomly" to describe both a cognitive process and a physical process. In the PostInterview, his range and reasoning both improved on "Six Samples" of the Large Jar, when he chose 48, 50, 58, 62, 68, 72, saying:

> GP: [Post Q1c] You want to pick around 60, kind of going a little extreme... I just picked around 60, and 10 or whatever [Waves his hands to show both sides of 60]. I mean, it could go anywhere...So I just picked more likely options.

GP knew the expected value in Post Q1c was 60 reds, but he didn't expect any of his "Six Samples" to actually be 60, just "around 60". He gives very reasonable choices. When he said "and 10 or whatever", he meant plus or minus approximately 10 on either side of 60. Even though he has the same "It Could Be Anything" kind of statement he made in the PreInterview, on the Post he stressed that he was picking

results that he felt were “more likely”.

The “Compare Lists” questions on the Pre and PostInterviews showed more of what GP *expected* and why. For example, on Q2 in the PreInterview, GP felt list (i) was “fine”, even though list (i) for six results of samples of the Small Jar has numbers that are only 6 and above (“7, 9, 7, 6, 8, 7”). Other subjects tended to notice that the entire list seemed high. List (ii) was most reasonable (“6, 7, 5,8, 5, 4”) and GP liked it because it had “less radical numbers”, which was the way he often referred to extreme results. List (iii) had all sixes, which GP did not like because “all in a row would be pretty unlikely”. GP thought results for multiple samples should be “random”, which frequently meant different or lacking a discernible pattern. Thus, he liked list (iv) – “2, 5, 4, 3, 6, 4” – because it was “kinda random, you know, not that many radicals in there”. He did not comment on List (iv) being low overall. He also liked list (v) – “3, 10, 9, 2, 1, 5” - saying:

GP: There’s the 1 and the 9, that’s pretty, you know...[High? Rare?] But I like that.
I: Isn’t that one most like the one that you put [On “Six Samples”] ?
GP: Yeah, I kinda...I did that because, I kinda wanted to be a little radical

GP did favor list (ii) overall, but it was clear from his responses that high or low results were fine with GP, and extreme values were not a concern, but he did like results to not all be the same.

On the similar Post Q2, list (i) for the Large Jar was also high (“72, 91, 74, 63, 81, 78”). GP checked off list (i) on Post Q2 as one of several lists he liked, saying the six results “just look like a bunch of random numbers, that were picked out of a jar.” Later in his response to Post Q2, GP eventually commented on the result of 91 being

"pretty rare". He still never commented on how list (i) was high overall. List (ii) on Post Q2 was the most reasonable choice ("61, 73, 56, 69, 59, 48"), and it was GP's favorite because they were "just pretty random numbers...they're all different, there's no rhythm to 'em." As in the PreInterview, on Post Q2 he did not like all the repeated values of list (iii), and again he didn't comment on how list (iv) was low overall. At the end of his consideration of Pots Q2's "Compare Lists", he commented on how list (iv) had a 34, but list (ii) had "less extreme numbers."

Towards the end of the PostInterview, on "Compare Lists for six samples of the spinner (Q11), again list (i) was high ("38, 43, 36, 26, 41, 33"). This time GP was more cautious, saying: "Um, I guess it's possible. The 43 and 41 is pretty high, but... Well, it's possible, I guess." Although he didn't comment on list (i) being high overall, he did focus on the extreme value of 43 for the spinner just as he had done for the extreme value of 91 for the Large Jar. Finally, on Post Q11 GP noticed how list (iv) was low overall ("15, 19, 11, 25, 21, 23"):

> GP: [Post Q11] I'd be surprised at this one too [List (iv)].
> I: Why?
> GP: Well, you have the 11...Yeah, these lower numbers, but...Possible. I mean...The highest one is, there's nothing over 25, so that's pretty unlikely.

For GP, the shift to the theme of possibilities and likelihoods showed a bit more hesitancy about accepting the highly improbably extremes shown in some of the lists on PostInterview Q2 and Q11.

GP mentioned experience as a reason for his expectations on both interviews. In "One Sample" of tossing the die on PreInterview Q9, GP was one of the two cases

who did not put all tens for the faces of the die. He put "7, 8, 9, 11, 12, 13" because:

> GP: Well, I knew it was going to be random, and so I first looked at the 60, and I said, well, these all have to add up to 60. So I divided by 6, and I said 10 each. And then I said, well, it's not going to be happening, 10 for each one, and so I just took 2 numbers and made it so they would equal 20, like 7 and 13 is twenty...8 and 12 is twenty, 9 and 11 is 20. And so I knew that would all add up to 60

GP again ties "random" to differences, he uses some part-to-whole reasoning, and also he knows that a uniform distribution is "not going to be happening." Thus, in considering "Who Cheated?" on Pre Q10, GP was quick to denounce Lee's results of all tens as unbelievable, saying "I look at it and go: Come on, Lee! There's no way that this happened!". GP also thought Lynn's results ("10, 11, 10, 10, 9, 10") "seemed too... Balanced. Too – Not as random, or something." For Pat's results ("2,15, 10, 28, 1, 4"), GP is explicit about relying on experience:

> GP: Pat...That's pretty, kinda believable, but ...[Takes his time thinking] um...Gonna...Too extreme, I guess...
> I: What tells you that?
> GP: Well, she only hit...with the 60 times, she only got one "5" ?
> I: Oh, yeah...
> GP: I mean, that's...You're going to get more "5"s than that, out of 60 rolls, you know?
> I: Okay
> GP: I've played board games, and I'd roll dice, and you get 5 more than that, you know

After the class interventions, GP referred to experience several times on the PostSurveys and in the PostInterview. For example, when considering different arguments for how twenty samples of the spinner might look on PostInterview Q12, GP said: "And that's when I would pull out the Phantom [Fathom] software and show 'em how this works." It was also clear from his comments in class that the activities

in sampling and probability, combined with the computer simulations, had made an impression on GP. He would point out results that had been obtained experientially as justification for what he expected.

Displaying: Whereas GP had some questionable expectations in other question involving sampling and probability on the PreInterview (and even on the PostInterview), I was surprised at how reasonable many of his ideas were in evaluating and comparing graphs. It seemed to me that he was more of a visual learner, attracted to graphs in the sense that he responded with much energy. For example, in PreInterview Q3 and Q4, when he was thinking about whether "Graph: 30" and "Graph: 300" were real or fake, he was quick to judge "Graph: 30" as fake, saying "I think they cheated". He thought there should be a wider range for "Graph: 30", but did think that the mode should be at 6 and the shape should resemble a "pyramid." He used "pyramid" several times in the interviews, often accompanied by holding his hands in an inverted "V". It seemed that "pyramid" was a way of connoting a bell-shaped distribution, and he justified his thinking of "Graph: 300" as real by saying:

> GP: [Pre Q4] I think this is more like the pyramid, what I would see. This looks more legit to me [Holds hands in inverted "V"]. It seems like it spreads out...you have a few extremes out here. and then it kinds goes up, where it is more likely in the middle here [Points to mode of 6]

GP appeals to all aspects of the distribution in one response: Average, range, shape, and spread.

On the similar Q3 and Q4 on the PostInterview, GP found it difficult to decide if "Graph: 30" was real or fake:

GP: [Post Q3] It's definitely possible. I can't see how you can say that this is...You look at it and go 'No, this is fake", you know? I just see that they're all...kind of gathering in the middle around 60. Anything is possible, you can't say 'Oh no, you guys did this wrong, you cheaters!' You know, I would say these are actual results: 'Good job, guys!' You can't prove that...they cheated. You can't.

However, on Pre Q3, GP had said confidently "I think they cheated", and I suspect that his softening of graph judgments might be linked to his sense of what was possible. Commenting on "Graph: 300" in the PostInterview, GP thought it showed actual results. His response included a focus on average and shape, and he also invoked experience as a justification:

GP: [Post Q4] The majority is over the sixty, kind of tapers off...That's usually the look of a large-number grab. You get more of that look. That's my experience.

When he said "over the sixty", he meant that the data was literally piled up above the mode of sixty.

A comparison of GP's responses to Q7 on both interviews indicates a situation where he showed more decisiveness as he "Compared Graphs" in the PostInterview than he had in the Pre. Looking at Graph 1 in the PreInterview, GP focused on the mode of 59:15, saying it was "really tall" and that "your eye usually goes to the tallest one." The mode was a visual attractor for GP. He then said "I think Graph 2 is more helpful", but as he explained his thinking, he started to do something that no other case did. GP started using his pencil to re-distribute data on the different graphs, trying to figure out how they compared to one another. He talked aloud as he shuffled data around, and seemed to come to an impasse about which graph was giving him more useful information. Then he said:

GP: [Pre Q7] I think this one [Graph 1] is easier...This one [Graph 2] gets a little confusing, you know. This one [Graph 1] is if you were going to talk about it, it'd be easier to do this one [Graph 1].

On the similar "Compare Graphs" question on the PostInterview, GP had an opposite opinion. That is, Graph 1, which had coarser rounding than Graph 2, was denounced as "totally misleading." He had some interesting comments about use of the average in either graph, saying:

GP: [Post Q7] I think the median and mean thing is kind of a tricky thing to use, in just weighing this one muffin. Because you're kinda compromising the weight, kinda thing, you know? You're just saying: You know, we didn't get one answer, so let's just...get the middle between the mistakes here...

I thought GP's ideas provided a basis for thinking about averages as a way of balancing out the variation ("mistakes") in the data. He went on to talk specifically about the rounding strategies used in generating both graphs:

GP: [Post Q7] Well, this [Graph 2] is more accurate because you're taking the less, the rounding – to the lowest quality, so you get a more accurate view of what you got. This [Graph 1] is more spread out, you know, less differences here [Graph 2] between the measurements, you can see.

GP uses "spread" to describe the range of Graph 1, which is wider than the range of Graph 2.

Just as GP talked about the mean and median being "tricky" in Post Q7, he also had some difficulty reconciling the identical averages on "MAX Wait-Times" with the differences in spread shown in the data sets (see Figure 30) in PreInterview Q8.

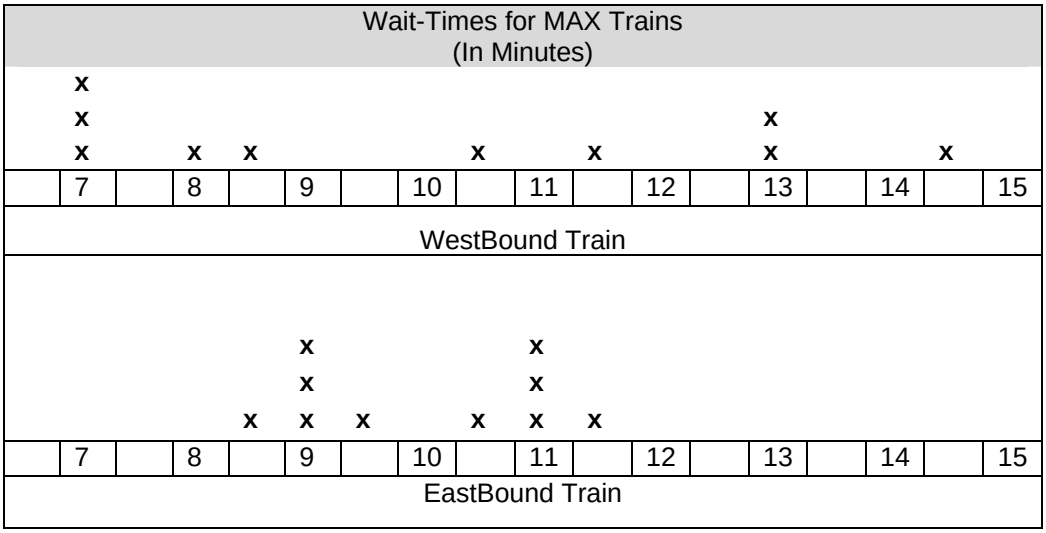

Figure 30 – PreInterview Q8 "MAX Wait-Times"

His initial impressions on PreInterview Q8 were:

> GP: [Pre Q8] Well, the Eastbound train seems to be more consistent. 'Cause their's seem to be more closer together [Pushing his hands together] This [Westbound] seems to be less consistent, since it's more spread out…[Drawing his hands apart]

His body language made me think that GP was primarily focusing on range, and when he considered the summary statistics, he seemed to experience the tension between centers and spread:

> GP: [Pre Q8] So in some ways it DOES balance out in the end, because these are balancing out as you see in these , these calculations. [Waves hands across the graphs, then points to summary stats] I'm going back and forth with these mean and median kinda things...
> I: Oh yeah? What do you mean?
> GP: Well, it's saying that they're the same in some ways, you know.
> I: Interesting
> GP: They have both the same numbers in the top and bottom [Same averages]. But then...they look different, down below [Different graphs]. So you're kinda going back and forth.

His eventual conclusion aligned with his initial impression, which he repeated, adding that the "Westbound train seems more of a...gamble."

GP correctly connected more variability with less reliability in the "MAX Wait-Times" question on the PreInterview, and he reasoned similarly on the Post Q9 about "Muffin Weights" (Figure 31).

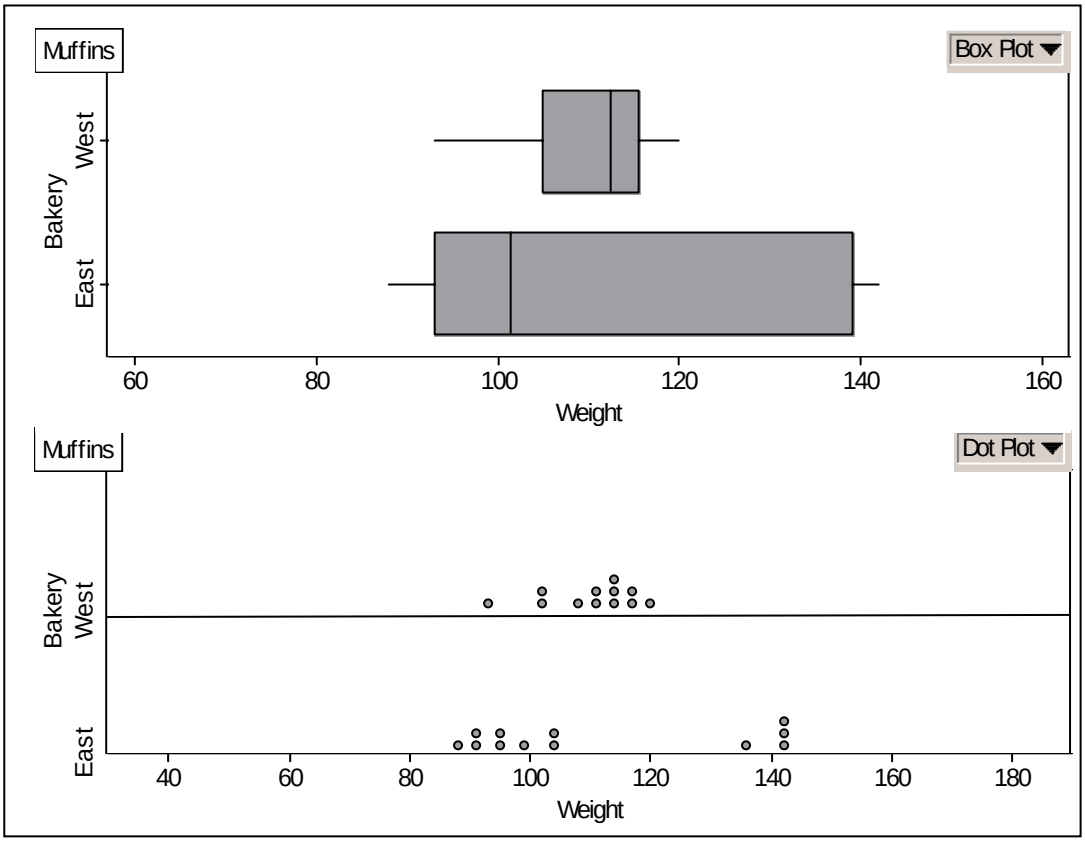

Figure 31 – PostInterview Q9 "Muffin Weights"

More importantly, the quality of his reasoning increased in sophistication in the PostInterview task:

GP: [Post Q9] Well, you see that the West End Bakery has a lot more consistency in their...the way they make their muffins.
I: How does the data show you that?
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 can 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 closer together.

GP's comments really focus on spread, especially the way he notices the "groupings" of the dotplot and the interquartile range of the boxplot. He reasons from both types of graphs for both bakeries in making his conclusion.

Interpreting: Many of the above responses have already showed how GP defined variation in terms of having difference, or being "random." Ranges were also a part of his definition, meaning a wider range corresponded with more "randomness." His response in PreInterview Q13 showed some of his ideas about variation as he reasoned about the "Likelier Graph?" :

GP: Uhhh, Group A is...more spread out.
I: Okay, what shows you that?
GP: You have some out here, like 13, 30...more randomness to it, I guess. This [Group B] is more bunched up. Probably Group A would be more expected.
I: Any reason for that?
GP: Just more random. I dunno

Later, GP noticed that "the middle [of Group A] is more random too," and he seemed to use the term "random" to describe spread as well as range. On the PostInterview he still used the term "random" in reference to variation he was noticing, but he also broadened his definition to include other terms. For instance, on the similar "Likelier Graph?" question on the PostInterview (Q13), he said about Class A and B's graphs:

GP: Well, Class A is a lot more compact, less range. Class B has a wider range, a lot more different variations...

GP's use of the word "random" featured prominently in an early PreInterview exchange in which he focused on the physical nature of the candy mixing. I asked him why he thought results for "Several Samples" of the Small Jar would probably not be the same each time:

GP: [Pre Q1b] Because, I mean, you pour in the different candies, and [they're] all mixed up, and you might grab in a different place or pick different ones, and...it's kind of random.
I: You're saying 'random'. What do you mean by that?
GP: Random being...You can grab from so many different places in the jar. So your hand can go this way, this way, this way [He mimics with his hands how he'd hold the jar and grab in different ways] this way, this way... And then, while you reach in, the candies move, all over the place, and so... Your hand creates randomness.

Since the above exchange was during the PreInterview, it was my first chance to see how really animated GP was. As far as his reasoning, GP clearly had a notion about the physical causes of variation. Most interestingly, he seemed to be saying that if only those candies would stop shifting around, and if one could grab the same way each time, then variation would be minimized.

He made fewer references to the nature of the candy mixing in the PostInterview than he had in the PreInterview. Instead, the bigger numbers of candies in the Large Jar of the PostInterview clearly caught GP's attention. On Post Q1a, when I asked about results for "One Sample", his first reaction was:

GP: [Post Q1a] Since there's more...you're more likely to get more of the, I don't know, extreme numbers...you know, the higher end and the very few, since there's more choices.

Thus, the sheer quantity of candies is, for GP, influential on expectation and variation of results. He repeated the theme concerning sample or population size a few times in the PostInterview. When comparing results of forty samples each from the Small and Large Jar on PostInterview Q5, GP was very clear that the Large Jar should have a wider range than the Small Jar:

GP: See, the thing is, this [Small Jar] is a wider range, it seems like, than this [Large Jar] … and this [Large Jar] should have the wider range.
I: 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, [Class B] should have a larger range than that [Class A], the smaller container.

Aside from the fact that GP's interpretation is contrary to what statistical theory suggests, the point is that GP clearly saw that the sizes of the jars influenced the expected ranges.

As a last example of GP's *interpretation* of variation, I knew from his PreSurvey responses that he thought more samples might increase the range of results. On the PostInterview, he added the idea that more samples also meant getting an average of results closer to the expected value:

GP: [Post Q1a "One Sample" of the Large Jar] The more you grab, the closer you'll get to 60 and 40 being...your 'grab', the more and more you grab...
GP: [Post Q10b "Compare Samples" of the spinner] It's all in the spin, you know. The more you spin it, though, you're gonna get closer to 25

His first of the above two comments makes it seems as if he was saying a single sample result will be closer to the expected value. However, I think based on some of the class activities where we kept a cumulative average of multiple samples, and based on some of GP's other comments, he meant that the average of all samples would move closer to the expected value.

In conclusion, GP developed some appreciation for how unlikely some extreme values were in sampling and probability situations, and he expects variation in results for multiple samples. My impression is that GP is a very visual and kinetic learner, which would be consistent with the way he instinctively seemed to relate to

graphs and also was concerned with physical causes of variation. In the classroom setting, GP was the only student who exploited the slight tactile difference in the chips we used for sampling. Thus, he was able to select all green chips, for example, and knew about skewing results due to physical manipulation. He reasoned fairly well with graphs, and incorporated new terminology and graph types into his PostInterview reasoning. He already attended to average and range in the PreInterview, and in the PostInterview he had increased attention to shape and spread. One concept that would be good to explore further with someone like GP is the effect of the sample and population size on expectation and variation, since GP showed some naive understanding of that effect on the PostInterview.

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 MET 1 the prior quarter with Steve. Although she was clear about not having had any prior courses in probability or statistics, her responses on the PreSurvey and PreInterview reflected her familiarity and comfort with mathematical ideas. 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.

Summary: EM *expected* some reasonable results in both interviews, but her sense of appropriate ranges and extreme values was not always consistent from Pre to Post. On the one hand, she predicted some reasonable ranges for "Six Samples" in both interviews. On the other hand, she had some poorer evaluations in "Comparing Lists" on the PostInterview than she had in the Pre. The biggest change for EM in this aspect was in her reliance on experience. Although she referred to her *own* instinct and experience on one PreInterview question, she made many references to *class* experience on the PostInterview. It was clear that the class interventions had made a significant impact on her reasoning.

When considering *displays* of variation, EM used more elements of the distribution in discussing graphs on the PostInterview than she had on the Pre. However, EM paid minimal attention to averages when evaluating graphs on both interviews, especially on both "MAX Wait-Times" (Pre Q8) and "Muffin Weights" (Post Q9). Instead, she talked in both interviews about how a train or bakery was consistent, and her reasoning focused mainly on the range.

There were two significant differences for EM in the *interpreting* aspect. The first difference was that while she hardly ever volunteered thoughts about the influence of the number of samples on expectation and variation in the PreInterview, she made reasonable several observations in the PostInterview. The second difference was how the PreSurvey and PreInterview picture I got of EM's perception of randomness and variation was that enough math or science could provide her with correct predictions, but she didn't convey any of those perceptions on the

PostInterview. A significant way in which EM was consistent on both interviews was in her attention to physical causes of variation.

Expecting: On PreInterview Q1a, EM guessed 6 reds for the results of "One Sample" from the Small Jar, and she reasoned proportionally. On the identical question from the PreSurvey, she had written "5 or 6", so she did have a sense that results from "One Sample" might not necessarily be the expected value. Her answers on the PostInterview "One Sample" questions were ranges:

> EM: [Post Q1a "One Sample" of the Large Jar] I think I would get, you know, somewhere in between 50 to 70 reds.
> EM: [Post Q10a "One Sample" of the spinner] I think it will land there somewhere close to 50% of the time...I think it will be probably between 40 and 60% of the time. Out of 50 spins, somewhere between – What would that be? Between 20 and 30 spins.

EM's adeptness at calculating 40% and 60% of 50 spins showed her proficiency in proportional reasoning, and she exhibited similar mathematical fluency in the PreSurvey and PreInterview. The more important point in the above two responses is EM's emphasis on range expectations in the PostInterview.

Her range for "Six Samples" from the Small Jar was adequate ("2, 5, 5, 6, 6, 8") on the PreInterview and also for the Large Jar on PostInterview ("54, 58, 60, 62, 65, 70"). Elsewhere on the Pre and PostInterview, when she had to list choices her ranges were also plausible. For instance, on Pre Q11 her results for "Six Samples" of sixty tosses of a die were "5, 8, 10, 12, 13, 15". On PostInterview Q10c, she predicted results for "Six Samples" of the spinner as "18, 20, 23, 24, 28, 32".

EM showed a peculiar inconsistency when "Comparing Lists" on the two interviews. In PreInterview Q2, she correctly thought list (i) – "7, 9, 7, 6, 8, 7" – was

too high and that list (iv) – "2, 5, 4, 3, 6, 4" – was too low. For list (v) – "3, 10, 9, 2, 1, 5" – she did not think the upper and lower extremes were likely. However, on PostInterview Q2, EM tended to favor high lists. She acknowledged that list (i) – "72, 91, 74, 63, 81,78" – was high overall and said she liked it anyway:

> EM: [Post Q2] I like the first one. Choice (i) has good variation, the numbers are also above 60, which I think is more likely to happen than below. I like that there are some 70s, I think the 91 is a little out there, but... Occasionally I think you are going to pull... [Something high]

When EM said list (i) had good variation, she meant that the numbers were all different from each other. I think EM had an unrealistic sense of just how unlikely 91 reds really is for sampling from the Large Jar. She seemed more cautious about the low list (iv) – "53, 41, 34, 60, 46, 52" – for which she thought both 34 reds and 41 reds were low. On the "Compare Lists" question for the spinner (Post Q11), EM again favored the high list (i) – "38, 43, 36, 26, 41, 33" – acknowledging that "it was definitely higher than the 50%." She cited class experience as a reason why she thought the results of list (i) were likely, and then she also went ahead and accepted the low list (iv) – "15, 19, 11, 25, 21, 23":

> EM: [Post Q11] I picked (iv) because I figured: If I'm gonna [go] high on number (i), I could see it going low, where maybe you would only get black 11 out of 50 times.

I was uncertain which class experience led EM to think that results might be generally high or low, since I don't recall any group presentations of such results. In class, we tended to aggregate results from multiple samples and those results were always on both sides of the expected value. However, the aggregate results encompassed more than six samples, so it could have happened that EM encountered runs of six samples

that reflected the high and low lists.

In any case, EM was certainly influenced by experience, and in the PreInterview she cited experience as a reason for why she knew she wouldn't get all tens for each face of the die in sixty tosses. At many other times in both interviews, EM stressed how results from multiple samples probably would not repeat each time:

> EM: [Pre Q2 "Compare Lists"] I don't think you're always going to pull 6
> EM: [Pre Q11 "Six Samples"] I'd be surprised if it came to be the exact same number as...before. I mean, yeah, It could happen, but I'd be surprised.
> EM: [Post Q1c "Six Samples"] I don't think they'll pull 60 every time
> EM: [Post Q10a "One Sample"] I don't think it will always be 50% of the time

On PreInterview Q9 ("One Sample" of the die toss) she put "8, 9, 10, 10, 11, 12" for each face, although she said "there's no reason not to get 10 of each." In fact, experience was EM's reason for not putting all tens, because later she said she was using "instinct, thinking about when I've played games, and how often sixes came up, and how often fours..." EM had naive reasoning about how often results might repeat when tossing a fair die, because earlier in Q9 she said:

> EM: [Pre Q9] Pretty often you're going to get, I don't know, every six or seven times [tosses of the fair die] I think you're going to roll the same number, again...

It is hard to imagine that EM actually based her response above on experience, and instead I suspect that proportional reasoning was influencing her thinking, based on her response above. Her informal experience also clearly influenced the list of numbers she provided in Q9, and she noted: "I just know that when I play games I usually don't get ones, so I made that one smaller."

On the PostInterview, EM made many comments about class experience. For example, on "One Sample" from the Large Jar, she reasoned that

> EM: [Post Q1a] From what we've done in class, when we pulled handfuls before, we can see that the numbers generally center around the same kind of percentage as there are red to yellow, so 60% red, 40% yellow, so somewhere between 50 and 70%

I liked how she included distributional language of how results would "center around" the expected value, and she appealed to a range of results. She based her response on what she saw in class, and she reasoned similarly in "Comparing Lists" on Post Q2:

> EM: [Post Q2, List (iii) – "60, 60, 60, 60, 60, 60"] Just from what we've done in class, I never pulled a number the same time, six times in a row
> EM: [Post Q2, 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. From what we saw in class, you know, it took I think something like 500 tries before we got so low a number.

EM's comment about the "500 tries" meant that she was recalling the computer simulation, since our hand-drawn samples usually totaled less than 300 for the entire class. For "Six Samples" of the spinner (Post Q10c), she knew "after having done it in class" that results would be "generally concentrated" around 25 blacks, which was a reasonable expectation. Experience was also her reason for why she liked the high list (i) on "Compare Lists" for the spinner (Post Q11), since she claimed that "I've seen it happen, so I liked it." Finally, using identical reasoning as she had for pulling all 60s from the Large Jar, she didn't like all 25s for the spinner, saying "I didn't pick 25, 25, 25...[List (iii)] – Although it's possible, I just haven't seen it happen, so I didn't pick that". I think it is good that the class experiences had left such an impression on EM, but I also think her thinking shows the dangers of relying too much

on experience. She often seemed to argue more on the basis of just what she had or had not seen, and less on the basis of what was more or less likely regardless of what she had seen.

Displaying: Throughout the interviews, surveys, and class interactions EM made it clear that she knew about averages and how to use them, but when talking about graphs EM tended to say more about other elements of the distribution besides the average. For example, in her evaluations of the "Graph: 30" and "Graph: 300" questions on both interviews (Q3 and Q4), EM paid more attention to range and spread than to average. She also paid minimal attention to average on the "Compare Graphs" questions (Q7), instead focusing more of her comments on the rounding procedures for each graph and spread of the data.

On PreInterview Q8, part of the interview script asked for her comparison of "MAX Wait-Times" when the averages were the same:

> I: [Pre Q8] A student in class argues that there really is no difference in the wait-times because the averages are the same. What would you say to this student...?
>
> EM: [Laughs] Ohhh. I know that the average says that, 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.

Aside from the time when I specifically directed her attention to the average, her other comments about the trains primarily concerned reliability and range:

> EM: 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. And nothing falling outside of that, so.

For EM, the term "variance" in her response above refers to the wider range on the Westbound train. On the PostInterview "Muffin Weights" Q9, the West End Bakery had the narrower range, and EM's reasoning was similar to that on the PreInterview:

> I: [Post Q9] How do you think these bakeries compare to one another, in terms of the muffins?
> EM: If I wanted to go to a bakery where I had a good sense of what I was going to get, they were more consistent in the weight of the muffin, I would go to the West bakery…
> I: Oh, why is that?
> EM: Because I can see from both the boxplot and the dotplot that they 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. 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.

As in the PreInterview, EM again paid most of her attention to the range. She reasoned both from the boxplot and the dotplot, considered spread as she talks about how data is "concentrated", and also included "risk" as a part of her decision-making process. She knew the dotplot gave more detailed information, and she explicitly tied the consistency of the West End Bakery to its narrower range:

> 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, and… That is more consistent, then.

A good example illustrating EM's overall increased sophistication in the PostInterview when reasoning about displays of variation is found in her answers to the "Likelier Graph?" questions. In PreInterview Q13, she mostly focused on ranges, and for Group B she noted:

EM: [Pre Q13] Group B is all...The number of times is all right in the middle of the graph, 17 to 23...Yeah, that would be exactly, really close to one-third of the time they landed on black. Which is what I guess would normally happen.

At first, EM seemed comfortable with the range of 17 to 23 on Group B, and she liked the spread of data being "close" to the expected value of 20 blacks, but she concluded that "Group A is more what I would expect." Her main reason seemed to be because she liked the wider range of Group A:

EM: There was the rare times when it [Group A] dropped less than 16, and way above 24. I can also see why that would happen as well. Where occasionally...the few that are way off the charts, you know, there's a 13...and then 30 times it landed on black.

She was clearly comfortable with the range for Group A, which had the graph reflecting actual data. Since she didn't explicitly bring up the average or the shape, I asked her how she felt about the fact that Group A only hit the expected value of 20 blacks one time out of twenty samples, and she countered: "But they got around 20: 21, 18, 19, somewhere in the... around one-third of the time."

Thus, EM's analysis on Pre Q13 was reasonable, but what I noticed in the similar question on the PostInterview was that her response took into account a better synthesis of the elements of the distribution. About Class A, she said:

EM: [Post Q13] For Class A, it's all – The numbers are ONLY between 20 – it looks like 22, and 28. Yeah, 22 and 28. And it's kind of, almost like a pyramid, with just a little drop off, after the 25, so it's shaped like a pyramid...it's all concentrated around the 50%

Thus, she included average, range, shape, and spread in her analysis of Class A, which she correctly thought was likelier to be fake: "This, to me, doesn't look right, that looks like somebody made that up." Her analysis of Class B was similarly rich in

detail:

> EM: Class B's are more spread out...your mode being 27, and you know, a lot – Several of the sets were within 24 and 28, and that's what I would guess. You have a 16, and you have a 34, and there is some variation in the numbers, and that seems to be more accurate because sometimes you CAN get as low as a 16, and sometimes you can get as high as 34...25 isn't the tallest number, so...heh heh

EM appeals to a subrange of 24 to 28 within which most of the data is clustered, and she also takes note of the extreme values, which are not unreasonable to her. She also doesn't mind having the mode be somewhere other than the expected value.

Interpreting: EM was the case who made the most references to class experience in the interviews, and she was also the case who made repeated mention of how she might answer if only she knew enough mathematics. However, she only talked about having enough evidence, or having a formula, on the PreInterview. It seems that her initial perception of variation was that she could make correct conclusions or inferences only with the proper knowledge. On the PreSurvey, for "Several Samples" and "Six Samples" (Q1b and Q1c on the PreSurvey were identical to those on the PreInterview) she wrote: "Sorry, but I don't know how to calculate these answers. I'm just going off instinct. No formula, just guessing." Here is a similar response from her PreInterview:

> EM: [Pre Q1c "Six Samples" of the Small Jar] 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

I included the entire transcript of her response to show how she wrestled with the question. Along with the effects on her perception – how she thought that maybe a "computation" would help her figure out results in the face of variation – EM also

exhibited the effects of variation on her decisions for “Six Samples”. On the die-tossing questions of the PreInterview, EM prefaced her comments about having played games by saying “I don’t have a calculation, but in my head, if I threw it, I think I’d see the same number at least once every seven times.” She used instinct and experience in considering results from tossing the die because she lacked “any scientific evidence for that...or mathematical evidence.”

Another major change for EM was in her sense of how the number of samples influenced expectations and variation. She only made one reference to the number of samples in the PreInterview, but made more then several such references on the PostInterview. Her thinking was that more samples would widen the range of results, and sometimes she used the inverse of this concept, meaning that less samples could have a narrower range. For example, when justifying her (reasonable) list of results for “Six Samples” of the Large Jar on Post Q1c, she stressed that “you’re only pulling six times.” A comparison of Pre and Post responses to the “Graph:30” and “Graph:300” questions exemplifies her attention to the number of samples. In the PreInterview, EM didn’t think “Graph:30” was real because:

> EM: [Pre Q3] I think, when you pull 30 times, you're going to have even more variety of times that you pull reds, and I can't believe that not once out of 30 times would they pull... no less than 5 reds.
> I: Ok, so it's the "no less than 5" that bothers you?
> EM: Yeah, it IS the "no less than 5" that bothers me...AND a little bit about the no more than 7. I think sometimes that you might pull 8 or 9, at least ONCE.

So, EM thought that 30 samples was enough to guarantee her wider range than 5 to 7 reds when sampling from the Small Jar. For samples from the Large Jar (Post Q3),

EM similarly thought about "Graph:30" that "maybe you would have a few over 70, or maybe one lower than 50, in thirty pulls." I commented that her own choices for "Six Samples" from the Large Jar (Post Q1c) had been between 50 and 70 red, and she countered that

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

She repeated her conviction about more samples having results that were more spread out later in analysing "Graph:300" on the PostInterview. Also, she considered 45 blacks for the spinner in Post Q11 "Compare Lists" to be too high for list (vi), but it could happen "if you did 5000 sets."

A final comparison of EM's thinking for the *interpreting* aspect concerns her attention to physical causes of variation on both interviews. For example, in sampling from the Small Jar in the PreInterview, consider her following responses:

> EM: [Pre Q1b "Several Samples"] Well, I mean, some yellows might get...extra yellows might get mixed in there. It's kind of the draw I guess, how many fall into your hand
> EM: [Pre Q1c "Six Samples"] Occasionally, maybe some yellows got pushed over to the side, so you'll pull more yellow
> EM: [Pre Q2 "Compare Lists"] In case, you know, some more yellow have gone into the batch in the jar...I think that in some places, you're not always gonna have a red/yellow red/yellow...In some places, they'll be yellows that have collected together.

EM created a very vivid picture of what she anticipated. She pictured possibly grabbing more yellows in her handful of ten candies because her hand might hit a pocket of overwhelmingly yellow candies. Her attention to physical causes was not limited to sampling, and in considering "Six Samples" of the die toss for Pre Q11, she

claimed:

> EM: [Pre Q11] I just don't think you're likely to get the same answer every time. There's no way you could do that unless you know how to drop the dice, or something.

Her implication was that someone could "know how to drop the dice" and thus get repeated results. In the PostInterview, she was concerned over whether the spinner was working properly on "Six Samples" of the spinner (Q10c), and her comments about the Large Jar sampling sounded very much like what she had said for the Small Jar:

> EM: [Post Q1b "Several Samples"] A big bunch of yellows might be there and that's where you reach. You're shaking it all around, but...that's where your hand goes, and so maybe you pulled some more yellows...

She mentioned physical causes less in the PostInterview than she had in the Pre, but it was clear that she was concerned about where those yellow candies were in the jar on both interviews.

In conclusion, class experience obviously had an effect on what EM expected. Although she frequently gave good ranges for results of multiple samples, she also allowed for some fairly unreasonable results because experience supposedly suggested to her that such results could occur or had actually occurred. Also, she knew that results for multiple samples would not necessarily be the same each time, but she had an interesting expectation of a pattern of results for the PreInterview die tossing question. She did not often mention average when considering displays of variation, but otherwise improved in her ability to talk and reason distributionally about graphs. Making decisions on the basis of where she perceived more reliability was important

to her on both interviews. She had increased attention to how more samples would broaden the range of results in the PostInterview, and on both interviews she seemed concerned about physical causes of variation. What stood out for me in EM's interpretations of variation was how she repeatedly seemed almost apologetic about not having the right math to figure things out on the PreSurvey and PostSurvey, but stopped reasoning along those lines in the PostInterview. The change in her thinking seemed to reflect 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 MET 2 course, she saw the variation in the outcomes of random events and wanted a formula. After multiple experiences with probability and statistics in class, she no longer mentioned 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 levity, such as when he mentioned bringing his "triple-beam balance" to the two bakeries in the "Muffin Weights" 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 MET 1 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 came 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."

Summary: JM was a strong proportional reasoner who shifted from having more emphasis on centers in his *expectations* in the PreInterview to having more emphasis on ranges in the Post. The biggest change for JM was that he seemed to contradict himself within the PreInterview but not within the Post. The two areas of contradiction for JM in the PreInterview concerned his sense of possibilities and likelihoods for extreme values and also for repeated values.

With *displays* of variation, JM demonstrated a very comprehensive and consistent ability to make sense of graphs on both interviews. However, even though he carefully analyzed every graph, on the questions that asked if the graphs were real

or made up JM seldom had confidence in making a decision. On such questions he tended to draw his own idea of what he thought the graphs should look really like. He was very consistent in using centers, ranges, shapes and spreads of distributions on both interviews.

There were two big shifts for JM in *interpreting* variation. First, he said much about physical causes of variation in the PreSurvey and PreInterview, and he said relatively little on the PostInterview. Secondly, on the PostInterview his ideas about the effect of more samples were more comprehensive than on the Pre.

Expecting: JM clearly knew how to calculate the expected value in figuring results for the "One Sample" questions, and in PreInterview Q1a his answer reflected his sense of proportion:

> JM: [Pre Q1a] I'd say, basically, six out of ten. There's a chance of six out ten.
> I: Why do you think that?
> JM: Well, because the ratio is 6 red for 4 yellow, for every ten there's 60%...

Similarly, on "One Sample" of the 60 tosses of the die in PreInterview Q9, he reasoned proportionally and focused on the expected value. He listed all tens for the faces of the die, and his justification was brief: "Well, it's one out of six." There were many other examples in the PreSurvey and PreInterview in which JM tended to emphasize a point estimate rather than talk about ranges.

In contrast, on the PostInterview all of his expectations were stated in terms of being close to or around the expected value. More importantly, his expectations on the PostInterview almost always included a range of possible results. He frequently

used the phrase “plus or minus” or some version of that idea to convey his range expectations. For example, compare his above response to “One Sample” of the Small Jar in the PreInterview to his response on the similar task with the Large Jar on the PostInterview:

> JM: [Post Q1a] Um, since the mix is 60% red...I’m gonna get close to that, maybe plus or minus...I’m going to say it’s going to be a good... Between, you know, 45 and 80

He later narrowed his range down to “45 to 75, somewhere around there.” He said he liked list (ii) – “61, 73, 56, 69, 59, 48” – because “it’s around 60, but it has a decent distribution of, looks like, 20% either way from the actual number”. In other words, JM liked the range around the expected value of 60 red. Also, in his comments about samples with the spinner on PostInterview Q10, for each part of the question JM emphasized ranges in explaining either *what* he expected or *why*:

> I: [Post Q10a] How many times, out of 50 [spins] do you think the arrow might land on black?
> JM: Well, approximately 50%, but it will be, you know, plus or minus, maybe 20% of that number – Somewhere in there
> I: [Post Q10b] Oh, the results on the second set, would be...
> JM: Yeah, I think [it’d be] fairly close in the sense that it’s gonna be around the...uh, 25 blacks, plus or minus that 10% or so...
> I: [Post Q10c] So, 21, 23, 25, 26, 27, 29! Why those numbers?
> JM: Well, they’re close to that 50 percentile that we’re looking for, plus or minus – I’m thinking, 10% or so. Actually, I’m a little high, aren’t I, with the 29? But still...

JM seemed fairly flexible with his ranges. At first (Q10a) he said plus or minus “maybe 20%”, and then backed down to plus or minus “10% or so” in Q10b and Q10c. What was significant to me was that JM clearly had a preference for range expectations in the PostInterview that was beyond what he indicated on the Pre. His

language of “plus or minus” mirrored what he and others in class had said when discussing predictions, particularly in talking about what was in the Unknown Mixture.

One area in which JM contradicted himself during the PreInterview concerned the possibility of extreme values. On Q1a, “One Sample” at the Small Jar, JM said “you could pick out 10 red. Or you could pick out 10 yellow.” Later in the PreInterview, on Q1c, he justified his prediction of “2, 3, 4, 5, 6, 7” for “Six Samples” by saying “you know, of course, you CAN pick out ten red, or you can pick out zero red.” By the time of “Graph: 300” in PreInterview Q4, JM backed away from his emphasis on the possibility of extreme values:

> JM: [Pre Q4] I can’t believe that there were, um, that you could...I don’t think you COULD pick up all 10 red, or all...or zero red. I think there would have to be some [yellow?]...I just think it’s impossible to pick out [all reds], if they’re mixed.

I did not point out JM’s inconsistency to him, but I found it interesting how the following pattern seemed to emerge with JM: When JM was asked to predict results, he was careful to mention how extreme outcomes were possible, yet when he was shown purported results (particularly in graphical form) he was skeptical of the extremes. On the PostInterview, JM still emphasized the possibility of extremes, but did not contradict himself. He said, for drawing “One Sample” from the Large Jar in Q1a, “I mean, it’s possible to get zero red, and it’s possible to get 100 red.” His most common way of talking about extreme values in the PostInterview included how those values were possible but unlikely. For example, in considering list (vi) – “30, 10, 90, 20, 60, 50” – JM commented that “when we go to extremes like that, they’re highly

unlikely and to have those...It's possible." I thought that JM had moved to a somewhat more balanced view about extremes in the PostInterview. Instead of swinging between polarized opinions of what could or could not happen, he had a better sense that some outcomes, while still possible, were "highly unlikely."

The other area in which JM contradicted himself during the PreInterview concerned the likelihood of repeated results. When I asked him in Q1b if he thought he'd get the same results every time for "Several Samples", JM was quick to stress "no, no...of course not." When I asked him why he thought results would not repeat, he said "well, because it's...it's impossible." Thus, it was clear that JM did not expect results to repeat for multiple samples, and on his own choices for "Six Samples" he did not repeat any values. However, by the time of PreInterview Q2, JM liked list (iii) – "6, 6, 6, 6, 6, 6" – saying initially that "Choice (iii)'s real good, because it could be 60% every time [laughs]." JM went on to wrestle aloud with the twin notions that getting all sixes was possible but unlikely. My main surprise for JM came in PreInterview Q9, when he listed all tens for the faces of the die in "One Sample" of sixty tosses, saying his choices were "not unreasonable." When JM considered Lee's supposed results of all tens in "Who Cheated" on Q10, JM said "I don't think, even if I rolled them 60 times, I would not get 10 numbers each." As I had done with DS, I pointed out how JM had listed all tens on Q9, to which he replied:

> JM: I think it's possible, likely, because it's one out of six times, but I don't think I could roll that and that'd actually happen. I said they're REASONABLE...

JM then went back and started to change his list of all tens on Q9, but then he resolved

to leave his list of all tens in place, and continued to defend his choice. His eventual argument was that

> JM: I think there's a greater chance of it coming up the ten, than maybe another number. And to pick ANOTHER number is, it'd be just as good as maybe picking ten. Since it's [the probability] 1 out of 6, I mean, I would pick 10.

JM was willing to stick with his own list of all tens, saying his list was "possible, but not probable", and clearly wrestled with himself over how likely he really thought it would be to see repeated results in multiple samples. In other tasks on the PreInterview, and all throughout the PostInterview, JM took a more moderate view of repeated values. That is, he included in his responses both the possibility and the unlikeliness of getting the same result each time for multiple samples.

Displaying: JM used all elements of the distribution at some point in both interviews as he deliberated the questions involving graphs. While he did not demonstrate any dramatic changes in reasoning about *displays* of variation from Pre to Post, he did add some sophistication to his discussion in the PostInterview as he invoked new ideas (such as the interquartile range) that he had gained in class.

To illustrate how JM reasoned about graphs, consider his response to PreInterview Q8 about the "MAX Wait-Times." At first, he pointed out how the two trains had the same means and medians. Then he said that there was a "big difference" between the trains because:

> JM: Even though we have 2 thirteens and a fourteen or a fifteen minute on the Westbound train, that's just 3 trains out of , what? Out of ten? Ten trains? So 30% of the trains take longer than 12 minutes on the Westbound train, and then , um, you know, it looks like 100% of the trains are under 12 minutes on the Eastbound train.

I thought JM's above response showed a reasonable attempt to compare the distribution of wait-times between the trains, and he later appealed to the shorter range of the Eastbound train in declaring it to be more reliable than the Westbound train:

> JM: Well, the Eastbound's are much more reliable. They go from 8 and a half to 11 and a half minutes. So you don't have this broad…You're going from 7 minutes to 14 and a half minutes on the Westbound train, so the chances of you waiting - shorter than the Eastbound train – are … it's only 30% of the time, most of the time you're going to wait equally or more, on the Westbound.

Thus, JM noted how the Eastbound train a narrower range than the Westbound train. He also calculated what percentage of the Westbound trains had longer and shorter wait-times than the Eastbound, which to me seemed to be JM's way of getting a sense of the relative spreads of the two data sets.

In the isomorphic "Muffin Weights" task on PostInterview Q9, JM reasoned similarly to the way he had in the PreInterview. Again, he first focused on the median, saying that "obviously the west End bakery produces, on average, a bigger muffin." Then, he shifted his attention to the range and spread of the data, using the boxplots as a basis of comparison. He rightly noted that "the interquartile range – 50% of all the muffins [in the East End bakery] – exceeds the whole range of the West End bakery." Because of their wider range, JM said "I'd probably be less confident of going to the East End bakery." He then made some astute observations about the three measures of center, which were all different between the two bakeries. For example, the West had a higher median than the East, but the East had a higher mean and a higher mode than the West. JM took note of how the data was distributed along the dotplots for the two

bakeries, and concluded:

> JM: 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.
> I: Oh, why is that?
> JM: Well, because there's just too many...the range is...too many low-weight and high-weight muffins. That really throws off our idea of the typical muffin.

I was pleased to hear that JM had some notion of the effect of variability on the mean in context of the "Muffin Weights" task, because a comparison of the merits of different measures of center had been a part of the class curriculum.

I'll use Q13 from both interviews – the "Likelier Graph?" task – to illustrate two tendencies that JM showed when comparing and evaluating graphs to see if one graph or the other was likelier to reflect actual data. One tendency was to not have much confidence in deciding whether graphs were real or fake, and the other tendency was to sketch on the interview script to show his own idea of what results should look like. He also showed these two tendencies in "Graph: 30" and "Graph: 300" (Q3 and Q4) on both interviews. In the PreInterview Q13, JM was quick to compare ranges in talking about which was the "Likelier Graph" :

> JM: [Pre Q13] Group A certainly has wider variables [Holds hands far apart], it's gone between 13 and 30. And Group B of course, is a much tighter distribution of black [Brings hands close together].

He then noted how Group A lacked any entries at the expected value of 20 blacks, while 20 blacks was the mode for Group B. For a time, he leaned towards thinking Group B's graph was real, but he pondered both graphs for awhile before saying:

> JM: These graphs have got me stumped in the sense that…I would think that there would be more… More 20s here [Group A]… I would like to meld both of the graphs.

JM decided that, instead of having confidence in either Group A or Group B being realistic graphs, he would draw on top of both graphs to make them look as he thought they should. He changed Group A by moving the mode to the expected value, and he narrowed the range. He changed Group B by reducing the height at the mode, and widening the range. In the end he had two graphs that looked roughly the same, and they were both shaped like smooth bell curves centered at the expected value of 20 blacks.

In the PostInterview Q13 "Likelier Graph?" task, JM reasoned much as he had in the PreInterview, but his language was slightly more descriptive. I asked him to compare the two graphs, and he said:

> JM: [Post Q13] Class A has that nice, nice shape that I was looking for [He draws an inverted-"V" shape on Class A], though it might not be evenly distributed. And Class B is just all over the place, with a mode of it looks like 27...And the fact that there were no, nothing below 22 here, on Class A, or above 28...it was just too tight, in 30 sets.

JM attended to average, range, shape, and spread in his response above, and eventually he concluded that both Class A and Class B "have a pretty good chance of being made-up." He then drew on both graphs. He expanded the "tight" range of Class A, and he shifted the mode on Class B to 25 while also narrowing Class B's range. Because he had mentioned Class B being "all over the place," I was not surprised to hear that he wanted data to be "more evenly distributed." It was clear that JM could use all elements of the distribution in making his evaluations and comparisons of graphs, but he still under-appreciated just how scattered data from on 30 samples could look in the PostInterview. That is, he saw the gaps along the axis for Class B

(the possible outcomes for which no results were attained), and he was skeptical. He also was still expecting to see the mode at the expected value, even for only 30 samples.

Interpreting: JM volunteered many more causes of variation in the PreInterview than in the Post. For instance, in PreInterview Q1c ("One Sample" of 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 ("Graph: 30") he expressed his thinking as follows:

> JM: You know, I guess I'd have to see how they fit into your hand. 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 [He means yellow]. Maybe? Something like that?

So too did JM stress causes in the spinner scenario of PreInterview Q12. Although Pre Q12 did not have an isomorphic counterpart in the PostInterview, JM's initial response focused on the mechanics of the spinner, not the content of the task. I'll 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.
>
> I: 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 side 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 ("Compare Samples"), and "getting into the rhythm of it [spinning]". However, JM volunteered much less about physical causes in the PostInterview than in the PreInterview.

Another change for JM was that he expressed better ideas in the PostInterview than in the Pre about what the influence of doing more samples would be on expectation and variation. On PreInterview Q1b, he indirectly appealed to the Law of Large Numbers when he commented about "Several Samples" that "I think on average, if you did it enough times, you probably average 6 reds". JM also suggested that more samples gave more chances to obtain the expected value in the "Compare Lists" question on PreInterview Q2, saying "in six tries, I would think that six reds would have to come up at least once or twice." More samples meant a broader range as well. In PreInterview Q3, JM felt that the 30 samples in "Graph: 30" would surely range beyond the 5 to 7 reds depicted in the graph. A significant idea JM expressed in

the PreInterview which was not expressed in the Post was how the ratio doesn't change regardless of the number of samples. On PreInterview Q9, JM argued in favor of his list of all tens for "One Sample" of sixty tosses of the die, reasoning that

> JM: You have one out of [six]...There's six sides. It's got to land on one of the six. And each one is, I guess, equal. So, after the first time, it's still one out of six. And the second time it's still one out of six. So...

He later re-iterated his emphasis on the unchanging ratio within the PreInterview.

On the PostInterview, JM didn't mention how the ratio is independent of the number of samples, but he repeated his earlier ideas and emphasized them more often. That is, more samples meant a widening range, more chances to actually attain the expected value, and a convergence of the cumulative average of results toward the expected value. Here are some examples of his responses:

> JM: [Post Q4a "Graph: 300"] 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...
> JM: [Post Q10a "One Sample"] If he does it enough times, he's going to be right at that number [6 reds]
> JM: [Post Q10a "One Sample"] With 50 spins...that's fairly good sampling, or...number of trials...that would approximate the theoretical probability

JM's reference to approximating the theoretical probability mirrored what we had discussed in class, and he was even more articulate when talking about the "Likelier Graph?" of PostInterview Q13:

> JM: [Post Q13] 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, chances are, they'll come pretty close if we do a fair number of sets.

JM also implied that graphs would have a better "look" with more samples, and about

"Graph: 300", he said that:

> JM: [Post Q4] I think that the more sampling that you do, the more close to that nice 60% distribution you're gonna get. If we did 3000 pulls, it would even look , you know, better…

I did not probe JM's thinking in the latter response, but I believe he was referring to having a smoother bell-shaped curve with increasing samples.

In conclusion, although JM gave more range expectations in the PostInterview than in the Pre, I could see how much he was influenced by centers in both interviews. For example, one reason he was unwilling to identify authentic graphs as such is because he thought the expected value was what should occur the most often, in both interviews. His expectations about ranges also seemed inconsistent at times. In PreInterview Q4 ("Graph: 300"), he thought the range was too wide for 300 pulls, and in the isomorphic task on PostInterview Q4, he thought the range was too narrow. In both situations, the graphs were authentic. JM was adept at using different elements of the distribution in discussing graphs, but he in both interviews he lacked confidence in deciding if graphs were real or made-up. Class experience seemed to help him identify unlikely graphs, but not to help him argue that a graph was in fact likely. I think that physical causes of variation remained an important issue for JM in both interviews, but it seemed more on his mind in the PreInterview than in the Post. He had many reasonable notions in both interviews about the influence of more samples, and clearly reflected ideas about the Law of Large Numbers which had been brought out in class.

The Case of SP

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 talked 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, telling me what would or wouldn't surprise her, for instance. Although she was an easy person to talk with, 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."

Summary: SP's PreInterview ideas about how "Anything Could Happen" and "You Can Never Know" reflect 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 think that when SP said that she

"can't guess", what she really meant is that she could not guess with the a priori assurance of being correct. In other words, she could never know ahead of time what the outcome would be. Since she believed in the PreInterview that uncertainty meant "Anything Could Happen", and because she could never know ahead of time what would happen, that why making a prediction was the same to her as "saying anything." Her sense of how "Anything Could Happen" explains why her ranges were so wide in many of the PreInterview questions, and yet it also explains why she gave all tens in Pre Q9 even though she didn't really think that outcome would happen.

SP *expected* results for multiple samples to be usually be different and not repeat, and she made explicit references to the underlying theoretical ratio in the PostInterview but not in the Pre. Instead of using "median" numbers and wide ranges to express what she expected in the PreInterview, she offered reasonable ranges that were appropriately centered around the expected value in the PostInterview. The distributional elements of range, shape, and spread were evident in her responses considering *displays* of variation in both interviews, and she included more of a focus on average in the PostInterview. In her *interpretations* of variation, during the PostInterview (but not in the Pre) SP volunteered some very reasonable ways that the number of samples might influence expectation. She also showed a major shift in her thinking, moving from the idea that "Anything Could Happen" in the PreInterview to the notion that some outcomes were likelier than others in the PostInterview. Her related PreInterview theme of not knowing gave way in the PostInterview to a theme suggesting that while you may not know for sure about a given outcome, you can still

make reasonable statements of expectation.

Expecting: SP established at many times throughout both interviews that she *expected* results for multiple samples to usually be different from one another. For example, in Pre Q1c ("Six Samples" of the Small Jar) she mentioned how she would "be more surprised if the same number kept showing up, as opposed to if it was just completely random." She acknowledged that repeated results were possible in "Comparing Lists" on Pre Q2, but maintained that the "6, 6, 6, 6, 6, 6" of list (iii) would cause her to "be VERY surprised." Even list (i) seemed "more unlikely" to SP, since list (i) contained three results of seven reds. Her responses in "Comparing Lists" on the PostInterview were similar to those on the Pre concerning repeated values. She liked list (ii) on Post Q11 because "there's not a lot of repetition". List (v) on Post Q11 – "24, 24, 25, 25, 26, 25" – was not favored by SP because of "too much repetition, you expect more variation."

The expectations that SP volunteered improved dramatically from Pre to PostInterviews. One area of improvement was how she gave appropriately wider ranges on the "One Sample" PostInterview questions. For example, in Pre Q1a for "One Sample" of the Small Jar, she said: "I guess instinctually I would say that it'd be somewhere in a median, like uh... 4, 5...just instinctually." In contrast, for Post Q1a she said that "One Sample" of the Large Jar should give her "somewhere between 50 and 70" red. Whereas she gave all tens for the "One Sample" of sixty tosses of the die on Pre Q9, for "One Sample" of the spinner on Post Q10a she expected "between 20 and 30" blacks. In the above examples, her expectations in the PreInterview are less

reasonable, than those given in the Post. A second area of improvement for SP was how the ranges she had for the "Six Samples" questions were appropriately narrower in the PostInterview than in the Pre. For "Six Samples" from the Small Jar (Pre Q1c), she listed "1, 2, 3, 4, 6, 8" , which is too wide. In the isomorphic question for the Large Jar (Post Q1c), she gave "49, 51, 55, 62, 65, 68", which is quite reasonable. Similarly, her choices for "Six Samples" of the die toss (Pre Q11) were "2, 5, 7, 1, 14, 20", again an unlikely list. SP had a much better list for "Six Samples" of the spinner in Post Q10c, "20, 23, 24, 26, 28, 29".

Another change for SP was that she included references to the theoretical ratio in the PostInterview but not in the Pre. For instance, she talked about expecting "median" numbers of 4 or 5 in "One Sample" of the Small Jar (Pre Q1a), and she repeated her preference for "median" numbers in response to a couple of other questions in the PreInterview:

> SP: [Pre Q1c] I just actually did like a median number, like 5...
> SP: [Pre Q2] And again, I don't know why I feel also comfortable with the median numbers, the 4, the 5, 6... for some reason. Yeah, like 4s, 5s, and 6, to me, is somewhere in the middle

It wasn't clear to me at that point in the PreInterview if SP even knew that the expected value for the Small Jar sampling was 6 reds. Even in the "One Sample" of sixty tosses of the die, she listed all tens but never explicitly said anything about the probability for any face being one out of six. Instead, she talked about giving all the faces "an even chance" and how she wanted to make sure her choices added up to 60. SP never articulated a single fraction or ratio anywhere in the PreInterview, showing what I think was an under-attention to the expected value. In the PostInterview, she

made it clear that she had considered the ratio in making her choices. For example, here is what she said in "Comparing Lists" on Post Q2:

> SP: There's also that 400 yellow in there, 600 red and 400 yellow, [and] the likelihood of just getting the 60 out of 100, which is like the perfect ratio or whatever, is very unlikely

An increased attention to the ratio helped SP to center her ranges in the PostInterview, such as when she had put "between 20 and 30" for "One Sample" of the spinner in Q10a, saying:

> 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] And uh, the between 20 and 30 would take into account the, the actual practice of spinning it…

The increased attention to expected values and improved sense of range that SP had in the PostInterview also allowed her to make better choices in the "Compare Lists" questions. In the PreInterview, she never commented on list (i) being generally too high, although she did feel the result of 9 reds was too extreme. Similarly, she liked list (iv) because of the "median" numbers, and never pointed out how the entire list was generally too low. She picked list (v) – "3, 10, 9, 2, 1, 5" – as her favorite because she liked the "huge...variety of numbers." She added that "Even though I said I would be surprised by 10... I feel like that [list (v)] covers a wider range . It has some high, and some really low." However, in the PostInterview, she correctly pointed out how list (i) was high overall and how list (iv) was low overall. She reasoned that

> SP: [Post Q2] 'Cause again – The perfect ratio's you would get 60 red, 40 yellow, and so...I guess you would want to go maybe 10 above or below that? Maybe more...for a variety of answers

At the end of her analysis, she picked list (ii) as her favorite on Post Q2, which was the most reasonable choice (and consistent with her own reasoning). List (ii) was also SP's favorite choice (and most reasonable) in Post Q11, the "Compare Lists" problem for the spinner. When I asked her why she liked list (ii), she said "they fall in that nice little range of mine, between 20 and 30, but with a few just going a little below and a little above, which I like." She also liked how list (ii) had no repeated values, and noted how the list had "a lot of variation." Thus, her final choices in "Comparing Lists" were better on the PostInterview than on the Pre.

Displaying: SP used range, shape, and spread in comparing and evaluating graphs during both interviews. The biggest change from Pre to Post was that she showed an increased attention to averages in the PostInterview. Also, while she correctly discerned real from fake graphs in both interviews, she seemed more confident of herself in the PostInterview.

I'll first illustrate her reasoning about *displays* of variation using her responses to the "Graph: 30" and "Graph: 300" questions on both interviews. In Pre Q3, at first SP spent some time wondering if she could have any confidence in knowing if the graph was real or fake. Eventually she thought that "Graph: 30" was made up because she did not "feel comfortable" with the shape of the graph, and we then had the following exchange:

SP: [Pre Q3] I like the wider range of things, I feel like that's more likely to happen. If, you know, to have it more random and this [graph] seems really less random.
I: Oh. What makes it seem less random to you?
SP: Because they're all 5, 6, and 7s. And three numbers in a row and...all clustered

SP thought “Graph: 300” was “more realistic”, and she liked it because “it’s spread out, there’s a little bit of everything, and then...most of them are somewhere in the middle...” She had a slight reservation about “how perfect” the graph was, and when I asked her what she meant she said:

> SP: [Pre Q4] As in, it, you know, it’s like this perfect curve. Whereas I don’t know if something’s randomly being chosen, that you can get this perfect curve. [Traces the shape with her finger]

She continued to emphasize shape as I asked her to compare “Graph: 30” to “Graph: 300”, saying that the former had “a more extreme curve” while the latter had “a more gentle, gradual curve.” I thought she reasoned well in the PreInterview, but what she added to her reasoning in the PostInterview was an explicit attention to the center of the distribution. Her entire response to Post Q3 is an excellent example of the overall improved caliber of communication:

> SP: [Post Q3] They seem like they could be the actual results.
> I: What convinces you, or what is your reasoning?
> SP: Because they fall into that sort of theory, that you’re going to have the most around 60, because of that perfect, that 60 to 40 sort of ratio, or 40 to 60, whichever. And so, you sort of that happening, and then they fall out in about a range of plus 10 , minus 10. So it makes sense. But they – 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.

She also mentioned how “Graph:30” was more “scattered” and that was why it was “not perfect” to her. She did a good job of synthesizing several elements of the distribution in her response, including an explicit mention of the average of 60 reds. SP also pointed out the mode of “Graph: 300” in Post Q4, and she commented on the graph’s shape, spread, and range in deciding the graph reflected genuine data:

SP: [Post Q4] But every once in a while it would happen that you would get, somewhere, like 48 or something like that, or a 73… And , yeah, it's starting to conform more to that bell-curvish, where you're getting mostly results from, like, 58 to 62 [The modal category] which you would think to happen, and again, it spreads out from there, along the range

I particularly liked the language she used when she pointed out similarities between "Graph: 30" and "Graph: 300" in the PostInterview, because it showed her attention to the way the data was distributed:

SP: Yeah, well, the bulk [of the data for "Graph: 30"] is still sort of in this little area here [She circles her finger around 60 red], and of course it's a little more scattered...And it does the same thing where it goes out almost at the same distance from that center 60.

SP conveyed a good sense of variation away from the mean in her latter response.

A second illustration of SP's reasoning about *displays* of variation comes from the "Likelier Graph?" questions on both interviews. Again, as in the "Graph: 30" and "Graph: 300" questions, SP demonstrated that she had some good ideas about graphs in the PreInterview which she improved upon in the Post. In PreInterview Q13, SP correctly identified Group A as likelier to be the authentic graph. She claimed that Group A had "greater variation" than Group B because Group A had a wider range:

SP: [Pre Q13] It's more spread out. It [Group A] goes from, the lowest is 13, and it goes up to 30. This one [Group B] is clustered within 17 to 23

In the isomorphic PostInterview Q13, she again correctly identified the authentic graph (Class B), and at first she used an argument based on shape which sounded very much like what she had said at the end of PreInterview Q4:

SP: [Post Q13] Class A is more like a drastic bell [Traces a bell curve], and this [Class B] is a more gentle bell [Traces broader bell curve]

SP went on to say how Class B had "more variation, is more spread out", while Class A was "more compact...and the range is really short." She explicitly mentioned the average (something she hadn't done in the PreInterview) when she pointed out that Class A had "not a lot of variation, and all sort of centered around that theoretical 25." SP also continued to refer to where the "bulk" of the data was for both Class A and for Class B, a term which she used to indicate where she saw data clustered.

Overall, the inclusion of the references to the average in the PostInterview was the biggest difference in SP's responses about *displays* of variation. She consistently reasoned well in the other questions involving graphs. For example, in the "Compare Graphs" questions for both interviews, SP felt that coarser rounding produced graphs that masked variation more than the graphs using finer rounding. In "MAX Wait-Times" and "Muffin Weights", she used both range and spread to help her identify the more consistent train or bakery. Throughout both interviews, she used many different descriptive terms suggestive of how data was distributed, such as "compact", "scattered", and "clustered."

Interpreting: One way that SP was fairly consistent in this aspect was how she referred to the number of candies in both interviews. In PreInterview Q1a, she said that "One Sample" of the Small Jar had "a slightly greater chance" of having more red "because there's more red than yellow." Later in the PreInterview, she repeated the same theme. In PostInterview, after she gave her range of 50 to 70 reds for "One Sample" of the Large Jar, she cautioned that she wouldn't expect "too many less,

because there're so many red in there." Later in the PostInterview she returned her focus to the number of candies used in sampling.

A significant change in how SP *interpreted* variation was that she mentioned the influence of doing more samples during the PostInterview but not during the Pre. Because I had asked questions on the PreSurvey and PostSurveys that directly invited thinking about the influence of more samples, it was clear that SP expected a wider range from an increased number of samples. However, she never volunteered any information in the PreInterview about the influence of more samples. On the PostInterview, however, she had several ideas. For example, in Post Q1a, she said "you expect a sort of, an average of 60, if you did many of these [samples]." I thought her response suggested the Law of Large Numbers, and reflected activities and discussions we had as a class. SP also thought that more samples gave more chances to attain extreme values, and she repeated this contention several times during the PostInterview. What follows are some of the different questions in which she addresses connects more samples to a widening range or more extremes:

SP: [Post Q1b "Several Samples"] But also [there'd be] some more extreme numbers, eventually

SP: [Post Q4 "Graph: 300"] They're going out a little bit even further, which which you would expect...With more pulls you would do, the more sort of outliers you would get, or the "unexpecteds" you would get...

SP: [Post Q11 "Compare Lists"] You're only spinning 50 times...But if you were spinning 100 times, maybe those numbers [extremes] would go further

The final idea SP had along this theme was that more samples influenced the shape of the graph. In "Graph: 30" she emphasized that "you only did 30 pulls, so it's going to

look a little bit more scattered." Since "Graph: 300" involved more samples, the graph "would become more...conformed to this perfect bell-curve, and that it would pull out just a little bit more." The ideas that more samples would move the cumulative average closer to the expected value, widen the overall range, and better reflect the shape of the underlying distribution were all ideas we had addressed as a class.

Another significant change was that SP repeatedly discussed in the PreInterview but not in the Post how she had difficulty guessing because "Anything Could Happen." The effects of variation on SP's perception and decisions were so pronounced in SP's PreInterview (and PreSurvey) responses that they seemed to dominate her thinking at times, and I'll highlight several examples. In PreInterview Q1a, her very first words for "One Sample" of the Small Jar were:

> SP: [Pre Q1a] My first instinct is just to say that, it could be any amount. You could have all, you could have none...[Then, later on:] But then, like, if I try to THINK about it, if I tried to think it out, then I'm thinking : It could be ANY amount, and I can't guess, you know.

She emphasized her view again in Q1b ("Several Samples"), saying that "it can be anything. Logically, that's what my brain is telling me, is it can be absolutely anything." I had already known about how SP thought results "Could be Anything" from her PreSurvey responses, but I then saw in the PreInterview how the effect on her decisions was that she did not want to make any guess at all. She continued in Pre Q1b to say:

> SP: [Pre Q1b] 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.

Coming at the beginning of the PreInterview, a snapshot of SP's thinking developed which portrayed her as having difficulty predicting results because "Anything Could Happen". By the time of Pre Q3, although she thought "Graph: 30" was fake, she was "also attracted to the 'We Have No Confidence' because we REALLY can never know, because it COULD happen, there's always the chance that it COULD happen." The effects of variation on her perception decisions were particularly relevant to her reasoning on the die-tossing questions of Pre Q9, Q10, and Q11. She reiterated the following view:

> I: [Pre Q9 "One Sample"] Why do you think those numbers [All tens] are reasonable?
> SP: Because if you're forced to guess...as I've been saying, that they could be anything. So I just gave them each an even chance...I can't guess. I have trouble making guesses

SP then went on to describe how she could get the same face of the die for each of her sixty tosses, but that she'd be surprised at that outcome because "it could be any of the numbers." She declared that "I'm just giving them each an even chance, because I guess...I can never know." On Pre Q10, in discussing "Who Cheated?", SP explained how her strategy in choosing all tens on Q9 was not motivated by what might really happen, but that "when you're making a guess, I just do it that way, because you can never really guess." She went on to stress how results "Could be Anything" later in her response to Q10, Q11 ("Six Samples" of the die toss), and Q13 ("Likelier Graph?"). Other subjects had expressed similar themes about not knowing what might happen, or what could happen, or how anything could happen, but no other subject was as outspoken on these themes as SP. Thus, I was surprised that in the

PostInterview, SP never expressed views about how "Anything Could Happen" or "You Can Never Know".

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. 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 MET 2 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 his own high school. Considering his own attitude at the start of MET 2, 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 corresponded well to the idea of variation from the mean, and his example of something that varies was "sea level."

Summary: Although RL clearly exhibited distributional reasoning prior to the class interventions, he had some contradictory *expectations* within both the PreSurvey

and the PreInterview. For example, on some questions RL wrote or talked about expecting to see variation in results, but on some other questions he said the expected value should occur repeatedly because it was the most likely outcome for a single sample. Due to the cognitive conflict induced by the sequencing of the die-tossing questions in the PreInterview, RL began a shift in his expectations that led to a more consistent appreciation for ranges by the end of the PostInterview. RL's reduced emphasis on centers from the Pre to the Post was also accompanied by a reduction in his frequent references to mathematical computation, and an increase in his focus on distributional reasoning.

RL seemed to misidentify real versus fake graphs for different reasons in considering *displays* of variation during both interviews, but overall I think he was relying on the Representative heuristic mentioned in Chapter 2. He liked "Graph: 30" in PreInterview Q3 because of the symmetry and the center, erroneously thinking that results for 30 samples would be a fair representative of the underlying distribution. He used the same kind of reasoning in considering the "Likelier Graph?" of PreInterview Q13. The graph for Group A he incorrectly labeled as fake because it was too "wild" and not as representative of the underlying distribution as Group B. In particular, RL thought that Group A had too many extremes. In both PreInterview Q3 and Q13, the small sample sizes used are not likely to yield graphs that are very representative of the population distribution, but RL did not seem to appreciate that fact. On the other hand, in PostInterview Q4, "Graph: 300" would be expected to give a reasonable

representation of the overall distribution. RL thought that the graph did not go wide enough, however, and was suspicious of the graph's authenticity.

In his *interpretations* of variation, RL reminded himself in the PreSurvey and PreInterview how reality was different from theoretical expectations, but he didn't often give voice to those thoughts in the PostInterview. He was the most outspoken subject in terms of the influence of the number of samples, especially regarding the influence on the shape of the distribution. He knew that overall ranges would expand with an increased number of samples, yet also expressed in the PostInterview how relative ranges would tighten. He seemed to focus more in the PostInterview on how the average of results of multiple samples should be the expected value.

Expecting: It is useful in RL's case to recall some of his original responses on the PreSurvey because those responses help establish RL's initial contradictions in terms of his anticipation variation versus his occasional over-emphasis of the expected value. He demonstrated distributional reasoning even in the PreSurvey as he considered both centers and spread in his responses. In explaining his choices of "4, 5, 6, 6, 7, 8" for "Six Samples" of the Small Jar on PreSurvey Q1c, RL said that "while 6 red candies remains the average outcome, variation is likely." Later, when reasoning about 50 trials at the Small Jar on PreSurvey Q3, RL claimed that "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 Samples" of the coin (PreSurvey Q7c), 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, even before the PreInterview.

However, RL was the *only* subject among the 27 who took the PreSurvey and put an unqualified "Yes" on Q1b when asked if results of several samples of the Small Jar would repeat. 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." He reasoned similarly in comparing samples of fifty flips of a coin on PreSurvey Q7b, saying that "in the absence of any change of approach, the results [25 heads] are most likely to be the same." This latter response is very telling about the thinking of RL and how the expected value sometimes dominated his reasoning at the beginning of the research.

In the first few questions of the PreInterview, RL still had an occasionally unreasonable heavy emphasis on the expected value. For example, he thought he'd get "probably 6" reds for "One Sample" of the Small Jar on Pre Q1a, and he then reasoned that for "Several Samples" 6 reds would continue to be the most likely result to occur. RL liked list (iii) on PreInterview Q2, and indicated that though rare, he would not be "too surprised" to see all six samples result in 6 reds each.

It was in PreInterview Q9 and Q10 that RL seemed to really begin a shift in his expectations. RL initially put all tens for "One Sample" of sixty tosses of the die in Pre Q9, saying that neither face of the die was "more likely or less likely than

another." Because I knew that RL had a strong math background and was likely to simply rely on proportional reasoning, I spent extra time with RL to make sure he understood that the intent of Q9 was for him to put what really thought might happen if we did the experiment of tossing the die sixty times in Steve's class. At two different times, RL repeated: "I think that's going to happen [All tens]." Again, I found RL's attraction to the expected value was at times a powerful influence on his expectations. But in Q10, when asked to evaluate Lee's list of all tens and Lynn's narrow list, RL was quick to point out "I don't believe they actually got that." He started to defend why it was reasonable for someone to predict all tens, but then he reflected on his own earlier thinking about "Six Samples" of the Small Jar. He had listed a reasonable "4, 5, 6, 6, 7, 8" for his "Six Samples" on Pre Q1c, which was the exact same list he had put on the PreSurvey (Q1 was identical on both the PreSurvey and PreInterview). He explained that he had originally thought of putting all sixes in his list for "Six Samples" of the Small Jar, but that he knew six reds wasn't "exclusive to all other possibilities." Thus, when ruminating over his list of all tens on Pre Q9 and reflecting on Lee's list of all tens in Pre Q10, he said:

> RL: That [All tens] would be the basis of my expectations. But it would be pretty funny, to see the likelihood matched so closely. Just like the other one [Flips back to Q1c, "Six Samples" of the Small Jar], just like this one [Q9]. Well, this was like, when I originally did this [Q1c], 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...

He then changed his Q9 list to "5, 8, 9, 11, 12, 15", and I noticed that he didn't even bother to include the expected value of 10 in his list. He was very articulate in explaining his change of mind:

RL: I'm changing my mind [on Q9] 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. This [All tens] is average only... there will always be a range of responses...but not every response will be 10. So, I think I was being limited in my consideration.

Pre Q9 and Q10 seemed to be watershed events for RL's thinking, in the sense that he really appeared to be engaged in some meta-cognition. He thought about how he had responded on earlier questions, and how repeated results of the expected value just didn't really make sense to him anymore. On PreInterview Q11, "Six Samples" of the die toss, RL listed "8, 9, 10, 10, 11, 12." He said that he didn't expect to see the same result for each of the six samples, and that he had "actually represented here a pretty limited range."

During the PostInterview, RL was clearly more appropriately attuned to range expectations than he had been in the PreSurvey or PreInterview. For "One Sample" of the Large Jar (Post Q1a), he initially said: "Well, I am inclined to estimate a range, rather than give an exact number." He then predicted "within a pretty wide range...I would say, even as low as oh, 40 to 80, even." After listing a reasonable "50, 55, 62, 65, 68, 70" for "Six Samples" of the Large Jar (Post Q1c), RL justified his choices by saying that "they are near the most likely value, but still wide enough to account for variation." Later, in PostInterview Q10a, RL thought that "One Sample" of the spinner would have a result "somewhere between 21 and 29...it's probably within that range." He felt in "Comparing Samples" for Post Q10b that results would "probably not" be exactly the same each time, but that results were "likely to fall in a same range, similar range" as he had given for "One Sample" on Post Q10a. He also gave a

reasonable list of “21, 22, 23, 27, 28, 29” for “Six Samples” of the spinner in Post Q10c, pointing out to me how “there’s no repeats, but...they’re similar, not identical.” When I commented on how his list did not include the expected value, he said that “25 is the theoretical expected result, but that’s not to say that that defines what happens.”

RL’s shift away from stressing the expected value went along with his marked decrease in expressing his mathematical calculations. While he stressed his mathematical computations in the PreInterview, he never gave voice to his calculations in the Post. In PreInterview Q1, RL wondered about the likelihood of getting all yellows (that is, no reds) in his sample of ten candies, and he calculated aloud:

> RL: Right. So, it is...if there is a 0.4 chance of pulling yellow, and then there’s 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 had considered 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 was unique in that no one of the other cases showed such a willingness to calculate to the same extent as RL. He even speculated on a rate of convergence in PreInterview Q2, noting:

> RL: Because if you’ve got...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 getting Red] it gets smaller more slowly.

There were many other places with both PreSurvey and PreInterview where RL offered fairly intense mathematical computations to analyze the situation, but not after the die-tossing questions in PreInterview. In the PostInterview, he still flavored his

explanations with relatively sophisticated mathematical terminology, such as when he wondered about finding the "inflection point" on a normal distribution of results, but he never articulated his calculations. Perhaps one reason for this shift is because his calculations were more useful to RL in finding centers, but not in estimating the variance that he came to expect.

RL's explanations in the PreSurvey and PreInterview often reflected distributional reasoning, but he was even more explicit about wanting a symmetric or skewed distribution in the PostInterview. For example, in PreInterview Q1a, when RL was talking about whether or not extremes were likely, he said:

> RL: Well, I mean that, 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.

I noticed how RL had used "bell curve" in his reasoning, even though there were no graphs provided for the question. It turned out that RL frequently envisioned what he thought the underlying distribution would look like, and occasionally he drew graphs on the interview script to help him make his point. In Pre Q1c, when predicting results for "Six Samples", RL started off saying "I'm going to use a bell curve, put 6 right at the middle here..." and then he drew a skewed bell curve and used it to help him make his choices. Later, in "Six Samples" of the die toss, RL mentioned that with each sample of sixty tosses "the whole distribution would be different." Thus, I could see how the distribution of results was important to RL even in the PreInterview, and I noticed how his lists on the "Six Samples" questions were symmetric around the mean

("4, 5, 6, 6, 7, 8" on Pre Q1c; "5, 8, 9, 11, 12, 15" on the amended Pre Q9; "8, 9, 10, 10, 11, 12" on Pre Q11). During the PostInterview, he continued to refer to distribution, often making clear his preference for a symmetric or a skewed distribution. In "Comparing Lists" on Post Q2, he liked list (ii) "because the graph of this distribution skews to the left, [so] I would think I'd see more pulls higher than 60 than less than 60." After he gave his range of 21 to 29 blacks for "One Sample" of the spinner in Post Q10a, he explained that

> RL: Because there's gonna be a, uh, symmetrical distribution, neither of these is more likely than the other...which is why I don't say, you know, 18 to 29...I'm going afar from 25 in either direction.

RL therefore explicitly detailed how he wanted variation on both sides of the mean. In the "Six Samples" of the spinner, he justified his Post Q10c list of "21, 22, 23, 27, 28, 29" by saying that he "did a pair, each equally far out from the mean in either direction." He continued to stress distribution in "Comparing Lists" for the spinner on Post Q11. Although he said "I expect to see a symmetric distribution" in rejecting list (i), the symmetrical list (v) was also rejected as unlikely, because RL said:

> 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.

The reason he favored list (ii) for Post Q11 was "because there's a range that seems legitimately wide, and it looks at first blush to be relatively symmetrical."

Displaying: RL attended to all elements of the distribution (center, range, shape, and spread) when considering *displays* of variation in both interviews, but he misidentified real graphs as fake and vice versa in each interview. In PreInterview

Q3, he thought “Graph: 30” was what he “would expect to see,” even though “Graph: 30” was really made up. Although he also expressed some surprise at the limited range for “Graph: 30”, he liked the fact that the mode was the expected value of 6 reds. He correctly thought “Graph: 300” showed actual results, and said that “the last graph here [Q3] was a very rough bell curve, [and] this [Q4] is much more similar to a bell curve” of the type he expected to see. For “Graph: 300”, he noted that “you see other things that are POSSIBLE, just relatively unlikely. You still see them come up, but just less often.” Thus, he commented on the shapes of both “Graph: 30” and “Graph: 300”, and also the ranges. He offered similar reasoning strategies in the PostInterview. However, while he correctly identified “Graph: 30” as real in Post Q3, he thought that “Graph: 300” might be fake (when it was not) in Post Q4. His concern about “Graph: 300” in Post Q4 was that he said “I would expect to see more, more extreme values...It’s kind of funny that there’s nothing outside that certain range.” I thought that in PreInterview Q3, at a time when RL was still fixated on the average, it was natural for him to accept the relatively tight range of “Graph:30” with minimal suspicion. He almost seemed to have over-compensated in his appreciation of range expectations in PostInterview Q4, since he thought that the range should be even wider than it already was.

A comparison of RL’s responses to the “Likelier Graphs?” questions in both interviews showcases his reasoning skills in evaluating and comparing graphs while also demonstrating an improvement in his final conclusions. In PreInterview Q13, RL thought that Group A was the fake graph and that Group B was the real graph, when

the opposite was true. About Group A, he said:

> RL: [Pre Q13] You have such a wide range...more outliers in Group A....Not only did someone get the very unlikely result of 13, somebody else got the very unlikely result of 30...Both of those [extreme values] are further out than everything else.

I noticed that RL never commented on how Group A only had one out of twenty results at the expected value of 20 blacks, but that his explanation hinged upon the wide range of Group A. His reasoning on Group B included a reference to center and spread:

> RL: Group B is tighter, and definitely holds to a center more...in Group B, you just didn't see that kind of variance [as in Group A]. In Group B, people did it, and no big surprises.

It did not surprise me that RL erroneously thought Group B was authentic, because the graph for Group B was almost symmetric about the mean, just as all of RL's lists were when he predicted results for multiple samples. In PostInterview Q13, RL correctly identified Class A as fake and Class B as real. He suspected that Class A had "underestimated, perhaps, the possibility of seeing less common values," meaning that he thought there should be more extremes in the graph for Class A. Furthermore, he said "the shape here [for Class A was] very tight, very narrow, not a lot of variation", and he thought the data was unnaturally grouped around the expected value of 25 blacks. Class A was "too neat" for RL to believe it came from genuine data. In Class B, on the other hand, RL saw "a lot more different values being represented," which he liked. He said he would have expected to see a graph like Class B's coming from real data because:

RL: [Post Q13] You still see in the middle there, between 24 and 27 or what have you, you see most values. And then a few on either side, kind of trickling out, or sprinkled to the sides.

I appreciated how RL attended to spread, noting that he pointed out "most values" in a central subrange of 24 to 27, and yet he also was comfortable with the few extremes shown in Class B.

RL used similar reasoning when comparing distributions on both the "MAX Wait-Times" question in the PreInterview and the "Muffin Weights" question on the PostInterview, but he increased his attention to subranges in the latter question. For PreInterview Q8, he thought the Eastbound train was more "reliable," and he pointed out how the Eastbound train had a shorter range than the Westbound. For the Westbound train, he made a point of stressing how a potential passenger wouldn't know what to expect for wait-times, because "there's a lot of variation...it's not a consistent pattern." When I asked him to comment on someone's argument that there was no difference in wait-times because the averages were the same, he said "the average does not tell all the story...they are not including the variation." He then affirmed his initial view that "the Eastbound is more predictable, and less variation." For the "Muffin Weights" of PostInterview Q9, RL named the West End bakery as being "a little more reliable." When I asked him why, he said that "the interquartile range is narrower, you can pretty much count on most muffins are gonna be within a certain range." I thought that RL used the boxplots in Post Q9 appropriately, and he went on to comment negatively about the wider range of the East End bakery. Just as he had wondered about how long a passenger might wait for the Westbound train in

the PreInterview, so too RL was concerned about the weight of the muffin he might get at the East End bakery:

> RL: The variation is SO much that, if I'm looking for reliability, if I wanna know what I'm gonna [get], 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...

He had a good observation in comparing the relative merits of boxplots versus dotplots, noting that "with the boxplot you're sacrificing information." When he wanted to see the specifics of where the data was grouped, he found the dotplots more helpful.

Interpreting: RL referred many times in the PreSurvey and PreInterview (but not as much in the PostInterview) to the way that theoretical probability was different from reality, and he seemed to use this theme to convince himself that results would not be the expected value each time. For example, in PreInterview Q1c RL had explained to me why he had initially put all 6s on the identical question earlier on the PreSurvey Q1c:

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

Later, in PreInterview Q3, even though RL thought "Graph: 30" was authentic, he wondered if one might see more extremes "just because weird things happen" in real life. Again, he used "living in the real world" as a reason for changing his list of all

tens in “One Sample” of the die toss in PreInterview Q9, mentioning our “world of imperfect scientific conditions.” He stressed that real-world “conditions are never identical”, and that was a reason he thought he would see variation in results. On the PreSurvey Q7b RL had implied that absent “any change of approach” in flipping coins, results would “most likely...be the same,” and yet by the end of the PreInterview he seemed to believe that variation was unavoidable. He did not say as much about the difference between reality versus theory in the PostInterview, perhaps because he had already convinced himself of the difference in the PreInterview and throughout the class interventions.

RL consistently described influences of the number of samples in both interviews, and he included many explicit references to the influence of doing *fewer* as well as *greater* numbers of samples. One characteristic that he frequently pointed out was how the number of samples affected the chances of getting extreme values. For example, in PreInterview Q4 RL had envisioned his own hypothetical jar containing 99 Red candies and 1 Yellow candy:

> RL: Just like the example I gave with the one yellow candy and the 99. You can pull out the yellow candy, it’s possible. It’s not going to happen that often, but if you do it enough times, sooner or later, it’s bound to happen.

RL was even more specific about the number of samples in PreInterview Q11, when he considered “Six Samples” of the die toss. He had listed “8, 9, 10, 10, 11, 12”, which we both thought had a “pretty limited range.” RL justified his choice as follows:

RL: But we're only talking about 6 people throwing, and when you've got 6, it's a pretty small sample size. So, chances are you're not going to see anything too goofy. You get a hundred people doing this, you're definitely going to see the extremes pop up more often.

I asked him what kind of range he might expect with one hundred samples, and he suggested the maximal range possible (0 to 60), saying "It can happen. It is unlikely." I thought RL had an under-appreciation of just how unlikely it is to get sixty 5s in sixty tosses of the fair die, but after the class interventions he had a better sense of how many samples it might take to attain extreme results. In considering samples from the Large Jar in PostInterview Q4, he talked about the chances of getting 0 red in his sample of 100:

RL: It's possible, it's a one in..., you know, trillion badillion, but if you do a hundred bunillion trillion pulls, you're gonna get a zero! And so, the more pulls you do, the more opportunity that exceptional event has of occurring.

He expressed the idea of more samples offering more chances to attain extreme results at many other times in the PostInterview, and was suspicious of unexpected results occurring with few samples. For example, in Post Q11 he declared list (i) for six samples of the spinner ("38, 43, 36, 26, 41, 33") to be high overall, and then he added:

RL: So, I'm a little bit suspicious, that having done so few spins, there's so many relatively unlikely [results]…With more sets, sure, I think again you're gonna start to see these more exceptional things happen, but you will also have seen many more expected results.

The last part of RL's previous response illustrates how RL also thought that more samples gave more chances to actually attain the expected value. In fact, he generally thought in both interviews that the average results of multiple samples should be the expected value. In other words, he did not expect the mean, median, or

mode to vary when doing more samples. Even in PreInterview Q9, when he finally talked about "living in the real world" and how results for sixty tosses of the die would not be all tens, he stressed that "if you're going to see a range, the average of that range will be 10." During the PostInterview, he repeatedly made it clear that results from multiple samples should have an average equal to the expected value. For example, in "One Sample" of Large Jar (Post Q1a), he thought that "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." Again, in "Several Samples" of the Large Jar (Post Q1b), RL thought that he'd "get more 60s than anything else", suggesting that the mode should be the expected value. He was even more explicit in PostInterview Q3 when he expressed how credible "Graph: 300" was, saying "we've got a mode of 60, which is what I would expect to see, so it looks believable."

Finally, in both interviews RL emphasized the influence of the number of samples on the shape of distribution of results, and he did so to a much greater extent than any other subject. During the PreInterview, for instance, as he compared "Graph: 30" and "Graph: 300" toward the end of his analysis in Q4, he noted that:

> RL: I expect to see a certain bell curve, given more trials. This was to so few trials [in "Graph: 30"], that it's not a very fleshed-out bell curve. Here [in "Graph: 300"] you start to see things fall into a pattern. If you did this [sampling] ten thousand [times], you'd probably have a really nice bell curve. So, I attribute the more bell curve-looking design to the number of trials.

I was impressed at RL's articulation in connecting the shape of the distribution to the number of samples, and he expounded on that connection even more frequently in the PostInterview. Even when no graph was present, as in "Several Samples" of the Large

Jar in PostInterview Q1b, RL's language reflected the shape of the underlying distribution:

> RL: [Post Q1b] 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. More, just – 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, actually.

At first I thought RL had concluded 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. His last comment in the PostInterview was a good exemplar of his view about the influence of more samples on the shape of distributions. After he considered the "Likelier Graph?" in Post Q13, he said in conclusion that "it's easy to see how more sets will start to normalize that distribution and approach the theoretical prediction."

In conclusion, RL had the broadest and strongest mathematical background of anyone else in the MET 2 class. He demonstrated distributional reasoning even in the PreSurvey as he considered both centers and spread in his responses. However, RL's thinking was occasionally over-influenced by the expected value in the PreSurvey and much of the PreInterview. After the die-tossing questions of the PreInterview, RL more consistently expressed his expectations in terms of ranges. Although he never once referred to what he had seen in class, I think that the class experiences

and his own self-reflection led to an improved sense of variation. His mathematical computations, which he articulated in the PreSurvey and PreInterview but not in the PostInterview, seemed to influence his choices throughout the research. Thus, RL carefully predicted results for multiple samples so that they were symmetrically distributed about the expected value. He reasoned about graphs using all aspects of the distribution (center, range, shape, and spread), and seemed to pay even more attention to relative subranges on the PostInterview. However, RL incorrectly identified real versus fake graphs on both interviews. Although he commented several times before the class interventions how reality was different from theory, after the interventions he stopped making those kinds of comments. Lastly, RL initially had a reasonable sense of the influence of the number of the number of samples on the distribution of results, and he demonstrated an even more extensive understanding of this theme in the PostInterview.

Cross-Case Comparisons

As mentioned previously, responses can be coded at multiple places within the framework, a possibility that arises when a response is longer and multi-faceted. From the interview transcripts, I took questions or portions of a question and considered each case's response through the evolving framework. Because some questions had multiple parts, there were often some substantial and lengthy responses by a case to a question. To illustrate what I did, consider Q2 ("Compare Lists" for the Small Jar) on the PreInterview. I asked subjects to pick the list(s) that they thought might be likely to occur as choices for six trials, and then to comment on all the lists. Then I asked

them which list they thought *best* described what might happen and explain why. Since there were five lists, naturally this interview question had the potential to elicit a considerable amount of dialogue in response.

In the cross-case analysis, I took each question or subquestion (such as Q1a, Q1b, Q1c, or Q2), and coded the aggregate response for each case. That is, I took everything the subject said on that question or subquestion and saw how the parts of the response fit into the framework. On Q2, for instance, I generated Table 13 to show me how the different cases' responses fit into the framework. I called such tables "CodeFrames" because I was coding responses in view of the framework. I made CodeFrames for every question on the PreInterview and PostInterview, including subquestions as I thought necessary or advantageous.

Table 13. *CodeFrame for PreInterview Q2 (Cross-Case Analysis)*

Framework	Description Within Themes	Subject (Case)					
1Ai	Should be on Both Sides of Exp. Val.	DS	EM		JM	RL	SP
1Aii	Won't be Exp. Val. Each Time	DS	EM	GP	JM		SP
1Aii	Shouldn't Repeat Values in General						SP
1Aiii	Should be in the MidRange		EM	GP			SP
1Aiii	Shouldn't be Too Many Highs (or Lows)		EM				SP
1Aiii	Should be Within Range Around Exp. Val.		EM		JM		SP
1Bi	Expected Value is Most Likely	DS			JM	RL	
1Bi	Extremes are Unlikely	DS	EM	GP	JM	RL	
1Bi	Extremes are Possible	DS				RL	SP
1Biii	Proportional Reasoning				JM	RL	
3Bii	Nature of the Candy Mixing		EM				
3Ci	Anything Can Happen				JM		
3Cii	Difficulty in Making a Choice			GP			
3Dii	Expected Value as an Average		EM				
3Dii	More Trials = More Variation	DS					

The CodeFrames give much information: the rows give different themes from the framework, or specific characteristics within the themes. The columns under the Subject (Case) heading show which cases were coded at the different places within the framework. For example, in Table 13, the CodeFrame for Q2 on the PreInterview shows how DS responded throughout Q2. Reading all the way down the Subject column for DS, we see that she had some part of her response address how more trials would give more variation. Moving across the row for "More Trials = More Variation", we see that no one else but DS included that theme as part of a response for Q2 on the PreInterview. On the other hand, Table 13 shows how five of the six cases all addressed three characteristics of themes in their responses to Q2. Results should be close to the expected value [1Ai], results won't be the expected value each time [1Ai], and extremes are unlikely [1Bi]. In this section I will summarize the rows (which represent dimensions, themes, or characteristics of themes from the framework) from PreInterview and PostInterview questions where all six cases had part of their response coded at that row.

As mentioned above, I had many CodeFrames for each of the interviews. The number of rows in each CodeFrame was inherently variable, depending on what the cases had to say. For instance, Table 13 has fifteen rows simply because that's how many dimensions, themes, or characteristics of themes occurred in the collective responses of the six cases. In some of the CodeFrames, there are matches among all six cases for certain rows, which I refer to as "Match 6 Rows". There were also "Match 5 Rows", meaning that exactly five of the six cases were coded along that row

(there are three such rows in Table 13). Table 14 shows the number of CodeFrames for the questions in each interview, the number of rows in each CodeFrame, and how many Match 5 or Match 6 Rows were in each CodeFrame.

Table 14. *CodeFrame Summary*

PreInterview				PostInterview			
Question Number	Rows in CodeFrame	Match 5 Rows	Match 6 Rows	Question Number	Rows in CodeFrame	Match 5 Rows	Match 6 Rows
Q1a	12	2		Q1a	11		1
Q1b	10			Q1b	10	1	
Q1c	11			Q1c	7	1	
Q2	17	3		Q2	14	2	3
Q3	10		1	Q3	12	2	
Q4a	12	1	1	Q4a	12		
Q4b	6			Q4b	8		
Q5	16		1	Q5	16		
Q6	4	2		Q6	6	2	1
Q7	18	2	2	Q7	17	3	1
Q8	15		2	Q8	13	3	1
Q9	7		1	Q9	16	1	
Q10	17	3	3	Q10a	8		2
Q11a	9	1	1	Q10b	6	1	1
Q11b	9		1	Q10c	10		1
Q12a	4		1	Q11	15	5	3
Q12b	6	1		Q12	14	1	3
Q12c	6			Q13ab	10	1	
Q13a	7			Q13c	14	1	1
Q13b	17						
20 Total	213 Total	15 Total	14 Total	19 Total	219 Total	24 Total	18 Total

I chose to show Match 5 or Match 6 Rows in Table 14 because those were the strongest two levels of agreement. As a percentage, the Match 5 or Match 6 Rows out of the total rows are [(15+14) / 213] = 13.6 % for the PreInterview and [(24 + 18) / 219] = 19.2 % for the PostInterview, suggesting slightly more agreement in the PostInterview. The fourteen Match 6 Rows in the PreInterview and the eighteen

Match 6 Rows in the PostInterview are summarized next, because they represent the most agreement.

PreInterview

Table 14 does not indicate the parts of the framework which garnered agreement, unless we see the actual rows from the different CodeFrames. Thus, I cut the rows out of each of the relevant CodeFrames and used the framework to re-organize the fourteen Match 6 Rows from the PreInterview. Table 15 shows exactly what the fourteen rows represented with respect to the framework.

Table 15. *Match 6 Rows from PreInterview CodeFrames*

Framework	Description Within Theme or Dimension	Question
1A	Riki: Really rolled It	Q10
1A	Yes, I'd be surprised if more Black than White in 3 spins	Q12a
1Aii	Repeated values could happen	Q11
1Aii	Their own choices are all different	Q11
1Bii	Probability arguments (chance, likelihoods)	Q9
1Bii	Extremes possible, but unlikely	Q10
1Biv	Lynn: Not enough variation	Q10
2Bi	Focus on mode in comparing graphs	Q7
2Bi	Comments on same summary statistics	Q8
2Bii	Noticing limited range: Only got three types of values	Q3
2C	Those are the real results shown in the graph	Q4
2C	Comfortable with average answer for "True duration of trip"	Q7
2Ci	Eastbound train: More consistent or reliable	Q8
2Cii	Engineer: Should use Graph 3	Q5

Table 15 shows where within the framework there was agreement among all six cases on the PreInterview. I'll discuss some of the areas of agreement from Table 15 in terms of the evolving framework.

Within the aspect of *expecting* variation, notice how all six cases thought Riki

was the student on Q10 (“Who Cheated?”) who really rolled the die. There is no theme within the dimension for that row, meaning that it is simply in the framework as [1A] because it shows specific expectations for that question. Most importantly, Riki did have the list which showed genuine data, and the fact that all six cases correctly identified Riki as having really rolled the die shows reasonable expectations on the part of the subjects. The theme concerning repeated values [1Aii] also had agreement in Q11 (“Six Samples” of sixty tosses of the die), and every one of the cases gave a list of predictions that contained distinct values. In other words, the lists given had no repeated values. I think that the subjects were particularly careful to choose all distinct values for their list of “Six Samples” in Q11 because the discussions of Q9 and Q10 tended to bring out strong reactions from the subjects about how six results of sixty would not occur. Thus, the subjects may have been over-compensating, thinking that if six results would not all be identical, then the six results would be all distinct. In explaining *why* they held their opinions, the cases agreed in Q10 (“Who Cheated?”) that extremes were possible but unlikely [1Bii], and that Lynn’s list did not exhibit enough variation [1Biv]. I was surprised that all the cases were suspicious of Lynn’s list, because I had thought some subjects might argue that “Anything Can Happen.” However, the common sense from the cases was that Lynn’s was too narrow, which is a reasonable assessment.

In *displaying* variation, all cases focused on the averages [2Bi] shown for the graphs in Q7 (“Compare Graphs”) and Q8 (“MAX Wait-Times”). Since averages are such a dominant part of traditional stochastics curricula and the media, it was no

surprise that my subjects' attention gravitated towards centers. I was encouraged to see that all subjects commented on the narrow range depicted in the graph on Q3 ("Graph: 30"), which did show fabricated data. However, even though they had a focus on the range [2Bii], not all the subjects identified the graph as being fake. In Q4 ("Graph: 300") the same specific conclusion [2C] for that question was made by all cases: Those were in fact real results shown in the graph. I also noticed that when making conclusions about the Eastbound train in Q8 ("MAX Wait-Times"), all cases had some emphasis on the consistency or reliability [2Ci] of the train. Finally, there was agreement that the Engineer of Q5 ("Car Brakes") should use Graph 3 in her report, a reasonable conclusion to make in the context of the question [2Ciii].

PostInterview

When I realized that there was agreement (Match 6 Rows) in the PreInterview for the *expecting* and *displaying* aspects but not for the *interpreting* aspect, I was curious to see how the eighteen Match 6 Rows for the PostInterview were organized according to the framework, and this organization is given in Table 16. Not only was there at least some agreement in the PostInterview on all three aspects, but the nature of the agreement represented an overall maturity of reasoning about variation. I'll comment more on this observation after discussing some of the areas of agreement on the PostInterview in terms of the framework.

Most of the agreement in PostInterview responses had to do with *expecting* variation. For example, on Q1a and Q10a ("One Sample" of the Large Jar and

spinner, respectively) each of the cases did not just put the expected value for their prediction, but rather they gave answers indicating an appreciation for variation.

Table 16. *Match 6 Rows from PostInterview CodeFrames*

Framework	Description Within Theme or Dimension	Question
1A	Gives a # Other than 60 or Range	Q1a
1A	Gives a # Other than 25 or a Range	Q10a
1A	Picks List (ii)	Q11
1Ai	Should be Close to the Expected Value	Q11
1Ai	Should be Close to the Expected Value	Q10b
1Aii	Their own choices are all different	Q10c
1Aii	Shouldn't repeat: Should be Different	Q2
1Aiii	Should have Variation or Range	Q2
1Bi	Extremes Unlikely	Q2
1Bi	Extremes Unlikely	Q11
1Bi	Extremes Possible	Q12
1Bi	No Guarantee of Getting Expected Value	Q12
1Biii	Proportional Reasoning	Q10a
2C	Class A : Likely Cheated	Q13
2Ciii	Rounding Affects Accuracy	Q7
2Ciii	More Detail in Histogram	Q8
3Bi	Operator Method or Perspective in Using the Scale	Q6
3Dii	Number of Spins Affects Amount of Variation	Q12

The six cases also all favored list (ii) on Q2 ("Compare Lists" for the Large Jar), which was the most reasonable choice. All cases gave responses that reflected the theme concerning the expected value [1Ai] in Q11 ("Compare Lists" for the spinner) and in Q10b ("Compare Samples" for the spinner). In particular, the cases' responses indicated that results should be close to the expected value. Further agreement for *what* was expected included the themes concerning repeated values [1Aii] and the idea that results should exhibit a range or some variation [1Aiii]. Regarding reasons *why* expectations were held, the language of possibilities and likelihoods [1Bi] was used by

all cases in response to several questions: Q2, Q11, and Q12 (“Compare Comments”). One key idea that seems commonly held is the notion that extreme values are possible but unlikely.

For *displaying* variation, everyone commented on Q13 (“Likelier Graph?”) that the graph for Class A was likelier to reflect made-up data, a correct conclusion [2C]. Concerning level of detail and usefulness of different types of graphs [2Ciii], there was consensus that less rounding led to a more accurate graph in Q7 (“Compare Graphs” for the muffins) and also that the histogram showed more detail in Q8 (“35 Muffins”). I noticed that there was no agreement for specific characteristics of themes within the dimension of evaluating and comparing graphs for the PostInterview, and I suspect one reason is that the questions offered more graph types than on the PreInterview, hence more opportunities emphasize themes in different ways.

There were two dimensions of agreement in *interpreting* variation. One dimension concerned causes of variation, in that all cases had some theme of operator error in using the scale for the repeated-measurement question involving the weight of a single muffin (Q6: “Causes: Muffin”). I considered the causes they listed as naturally occurring causes [3Bi] because they did not include a deliberate, subversive attempt to introduce variation, but were the kinds of variation that one would reasonably expect to find among different people attempting to discern a measurement. Finally, in Q12 (“Compare Comments”), everyone had some element of their response that connected the number of trials or spins with the resulting variation [3Dii].

In summary, the use of the CodeFrames in the cross-case analysis reveals some overall trends, most notably the closer agreement in *expecting* variation in the PostInterview. For example, on the PreInterview there were many predictions for "One Sample" questions that were just the expected value. However, on the PostInterview ranges were given for predictions, or values that were explained as being "near" to the expected value. Also, on the PreInterview, some cases did not access proportional reasoning, while others seemed overwhelmingly influenced by theoretical predictions. On the PostInterview they all used proportional reasoning but no one claimed that the theoretically expected value should always be the outcome. There were also some uniformly reasonable conclusions made regarding *displays* of variation, as well as attention given to average and range when evaluating and comparing graphs. Finally, in the PostInterview there was total agreement about plausible *interpretations* of variation, namely the cause of variation in Q6 ("Causes: Muffin") and the influence of more trials on results in Q12 ("Compare Comments").