STUDENT WORK AND STUDENT THINKING: AN INVALUABLE SOURCE FOR TEACHING AND RESEARCH

J. Michael Shaughnessy
Portland State University, United States
mikesh@pdx.edu

What can we learn from our students' thinking? This talk will share examples of student work (in both written and video format) as they wrestle with such concepts as variability, randomness, and sampling distributions. The examples of student work come from both written and interview data gathered in a research and teaching project that has been investigating the development of students' conceptions of statistics, particularly their conceptions of variability. The students are middle school and secondary school students from six different schools in the Northwest part of the United States. The concept of a sampling distribution is one of the biggest ideas in all of statistics, and provides one of the connections between probability models and inferential statistics. When a sampling distribution is created either from repeated trials or from repeated samples from a population, both randomness and variability play major roles. We assume the sampling distribution was generated under conditions of randomness, and we expect to learn something about variability from the distribution we obtain. How do our students think about randomness, variability, and sampling distributions, and what can we learn from their thinking that will help our teaching of statistics?

BACKGROUND

In recent years there has been increased attention within both the research and teaching communities of mathematics education to mining student work and student thinking to obtain clues about how students develop and construct their mathematical knowledge. Teachers and researchers in statistics education have also begun to look more closely at student work and student thinking on statistics tasks, in order to gain better insights into what their students know about statistical concepts, and how they come to know it. For example, Mokros and Russell (1995) interviewed Grade 4-6 students using tasks designed to explore a variety of students' conceptions of centers. They were able to identify student thinking that typified average as mode, as algorithm, as 'reasonable', as midpoint, and as point of balance among the students they interviewed. While researching students' informal reasoning strategies when comparing data sets, Watson and Moritz (1999) interviewed students in Grades 3 – 9, and found that students made inferences and decisions based on a variety of approaches, including purely visual strategies, modal clumps of data, additive strategies (e.g., this group has 'more' than this group) and ultimately proportional reasoning strategies. Reading and Shaughnessy (2004) discuss student thinking about variation during interviews involving a sampling distribution task. They categorized students' predicted sampling distributions as wide, narrow, or reasonable (for variation) and, low, reasonable, or high (for centers), and provided examples of students' reasoning for their predictions for the sampling distributions. Bakker and Gravemeijer (2004) analyzed student work and student discourse as students made comparisons between two distributions of data, and subsequently used the students' thinking to amend and build teaching episodes on comparing distributions.

These are just a few examples from research and teaching over the past decade which demonstrate increased attention to students' thinking for both research and curriculum development. One of the benefits of documenting and analyzing student thinking on statistical tasks is that researchers and teachers often learn much more than they had originally attempted to discern. Unlike survey tasks or traditional tests, much more of the details of student thinking can be exposed when we ask students to give reasons for their answers, or to make conjectures about what they think, and why they think it. In fact, when student reasoning from task-based interviews is coupled with large survey results on identical or similar tasks, teachers and researchers enter a data-rich environment that provides both the necessary detail as well as the sample size numbers to obtain both valid and reliable indicators of student thinking.

OUTLINE

In this plenary talk I will share some excerpts of student thinking and reasoning on three statistical tasks that were originally designed to probe for students' understandings of variation and variability. Task based interviews on data sets presented in both graphical and tabular form were conducted with school mathematics students in Grades 6 – 12. Students' responses to these tasks provide a rich texture of their thinking about aspects of variation (Wild and Pfannkuch, 1999), such as acknowledging and attending to variability, suggesting causes of variability, and appealing to variability when comparing data sets. However, as researchers we learned far more than just what students were thinking about variability on these task. Students' context knowledge, or lack of it, played a major role in their thinking about the data sets. The effects of their own personal experiences on their reasoning; their ability to read graphs; their integration of centers, spreads, and shapes of data; their beliefs about chance and sampling—all of these were also uncovered during our interviews with the students. Although the students were middle and secondary level students in school mathematics courses, we have found that their thinking and reasoning is not unlike undergraduate statistics students that have been interviewed and surveyed on similar tasks (Ciancetta and Noll, 2006). The results of our interviews suggest that there are levels of sophistication in students' thinking on statistical tasks through which all students grow as their thinking about data sets matures. Furthermore, the sophistication of students' thinking on these tasks is not simply age related, but also depends on students having frequent experiences with data analysis throughout their school years.

PROCEDURES AND STUDENTS

The students in our research project came from six schools, two urban, three suburban, and one rural. The students were in ten classes, three middle school and seven secondary school classes, taught by mathematics teachers who had agreed to work closely with our project over a three-year period. These teachers acted as co-researchers, and our research team in kind acted as co-teachers, over the duration of the project. We conducted three week long classroom 'teaching episodes' during the project, one on sampling distributions, one on repeated measurements, and one on mining the data within a rich multivariate context about fast foods. Over 236 students in these classes were surveyed both before and after the teaching episodes on tasks similar to those used in the classroom teaching episodes. These students had had a variety of different kinds of experiences with stochastics. The middle school classes had previous experiences with reading and making graphical displays of data, and conducting probability simulations. The secondary classes spanned the continuum of first to fourth years of secondary school. The secondary mathematics curricula of our students went from first year algebra with little or no statistics experiences, to courses with imbedded experiences collecting and graphing data, all the way to several AP (Advanced Placement) statistics classes. The tasks and student responses discussed in this plenary are excerpts from the third and final interview with these students which was conducted near the end of the second year of our project, after all the teaching episodes had been completed. In each of the six research classes four students—two boys and two girls—were randomly chosen for interviews on these tasks. The other four classes served as comparison classes that were surveyed only and did not experience the classroom teaching episodes. The students were ordinarily interviewed during math class time, and each of the task-based interviews took about 45 minutes to complete.

THE INTERVIEW TASKS

The tasks in the final interview touched on some of the major issues that had surfaced during each of the three teaching episodes. I will discuss student responses from three particular tasks in our research. Each of these three tasks was intended to get students to attend to various aspects of variability (the propensity for things to vary or be different) and/or variation (ways of measuring of variability). This distinction between the words variability and variation has been discussed in Reading and Shaughnessy (2004). After presenting the tasks I'll share examples of student work and reasoning about each one. The reader is encouraged to think for themselves about each of these tasks before reading the student responses. How might your students respond to these tasks? How would you respond?

Task 1

In Task 1, we asked students "What do you notice? What do you wonder about?" to see what sort of unprompted thinking might arise about this graph of fruit juice consumption over time. Nemiriovsky (1996) has suggested that graphs over time are the best place to start to introduce and explore variability with students. Parts b) and c) of Task 1 attempt to probe for students' understanding of global trends, and for potential sources or causes of variability in fruit juice consumption. This task is similar to some tasks that students explored in the third teaching episode in our research. The data for Task 1 were obtained from the U.S. Department of Agriculture (2004).

Task 1: *Fruit Juice graph of consumption over time.* (in gallons per person per year).

a) What do you *notice* in the graph? Is there anything you *wonder* about?
b) Can you describe how consumption of fruit juice varied over time?
c) Can you think of some reasons that might explain the variability?

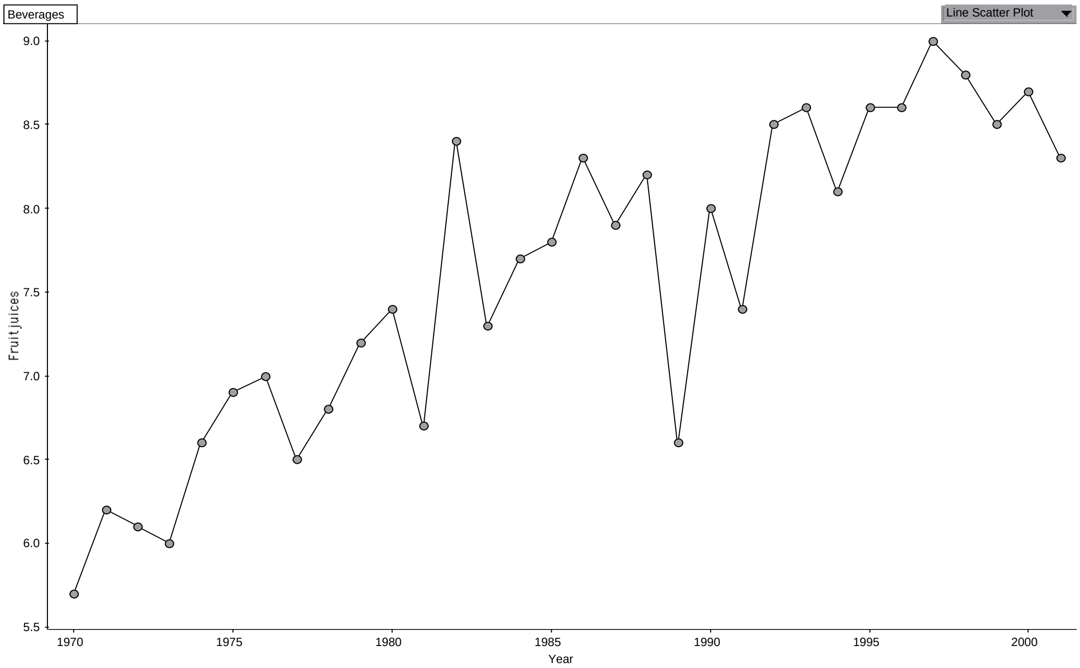

Figure 1: Fruit Juice graph of U.S. consumption in gallons per person per year

In addition to unprompted "notices" and "wonders," this task revealed students strategies for reading such graphs. Some students attended to individual data values, others discussed overall trends, and still other students noted both local and global aspects of the graph. Reasons for the variability in the graph went from 'don't know' to wild non-contextual speculations, to a focus on contextual issues such as marketing, drink competition, and fruit growing conditions.

Task 2

Task 2 mirrors data sets that students collected during the second teaching episode in which the students in the research classes collected repeated measurements of one students' body parts such as arm-span and head circumference. Once again we asked students "What do you notice" and "What do you wonder about?"

Task 2: *Repeated Measures.* Two different classes of students were gathering measurement data. Everyone in each of the classes measured Stephanie's foot length then graphed their data. Here are the data for the two classes.

Class A: Foot Lengths for Stephanie-- measured to nearest half-centimeter

												x						
												x						
										x		x						
										x		x						
						x			x	x		x						
						x			x	x	x	x		x				
	x	x		x		x			x	x	x	x	x	x		x		
17		18		19		20		21		22		23		24		25		26

Class B: Foot Lengths for Stephanie-- measured to nearest half-centimeter

										x								
										x								
										x								
										x								
										x								
										x		x						
										x		x						
										x	x	x						
										x	x	x						
								x		x	x	x	x					
								x	x	x	x	x	x	x				
17		18		19		20		21		22		23		24		25		26

a) What do you notice in the data? What do you wonder about?

b) Each student had his/her foot length measured by one of the other students in the class. Then a graph of all the foot lengths was made for each class. What might you expect the graphs to look like for each class of the two classes? Why?

We were interested in whether students would address variation between the two samples in Task 2 as well as variation within each sample. How would they account for the variation in measurements of one person's foot length? Would they predict wider spreads for class-wide foot measurements? Would they recognize the possible measurement reliability issue for Class A when the classes measured everyone's foot?

Task 3

One of the principle goals of our research was to investigate students' attention to and understanding of variability in the context of repeated sampling from a known population. In particular we were interested in what students would predict for repeated sample proportions, such as in Task 3. Would students predict wide spreads or narrow spreads? What attention would they give to potential outliers, too much or too little? Would they predict that the same result should happen every time, because 'theory' says that should happen?

Often in school mathematics tests students are asked to predict the outcome of a single sample pulled from a known population. An example of this occurred on the 1996 National Assessment of Educational Progress (NAEP) as discussed by Zawojewski and Shaughnessy (2000). Such questions are meant to assess whether students reason proportionally, from the population to the sample. For example, for a single pull in Task 3 below, a student might predict 65 reds in a sample of 100 candies since that mirrors the population proportion. However, this type of question attends only to the expected 'center' of a distribution. It is very similar to just asking a student "What is the probability that we would pull a red candy?" The overemphasis in school mathematics on measures of center as opposed to measures of variability has been well documented (Shaughnessy *et al*., 1999). Thus in our research we developed teaching episodes in which students first predicted sampling distributions, and then created them by pulling repeated samples from large buckets of colored chips and graphing the results.

Task 3: *Repeated Sampling I.* Suppose that you have a very big container with *1000* candies in it, 650 are red, and 350 are yellow. The candies are all mixed up in the container. Students scoop out 100 candies at a time, count the number of reds, and then replace and remix the candies.

a) What would you predict for the numbers of reds in pulls of 100 candies by 6 students?

______ ______ ______

______ ______ ______

Why do you think this?

b) Suppose that 50 students each pulled out 100 candies, from the bowl, wrote down the number of reds, put them back, mixed them up. Of the 50 students, how many of them do you think would get:

0 to10 reds?________	51 to 60 reds?________
11 to 20 reds?________	61 to 70 reds?________
21 to 30 reds?________	71 to 80 reds?________
31 to 40 reds?________	81 to 90 reds?________
41 to 50 reds?________	91 to 100 reds?_______

Why do you think this? Explain your choices.

Since a large number of the students in our research classes seemed to believe that unlikely outcomes would occur on this task, such as 0 to 10 reds in a handful of 100, we devised a follow-up question to explore students' conjectures about an unlikely event, and we ran computer simulations to create sampling distributions to test their conjectures:

Task 3: *Repeated Sampling II***.** Suppose that we have a mixture of 100 candies, 65 red and 35 yellow. We pull out 30 samples of 10 candies, replacing and remixing after each sample of 10.

a) How many times out of 30 tries do you think we will get 'no reds' in our samples of 10*?* Why?

b) What if we did it again, how many times do you think "no reds" will occur in 30 tries this time*?* How are you thinking about it?

c) How many samples of 10 do you think it would be necessary to pull before we obtained at least one sample of 0 reds*?*

This third task has helped us to document and develop a lattice of levels of student thinking about sampling problems. Some students think about sampling only in terms of absolute frequencies, purely additively. Others identify the relation between population proportions and sample proportions, but may fixate on centers and modal predictions with little attention to variability in the sampling results around a center value. Finally, there are students who we classified as *distributional* reasoners, who explicitly attended to both centers and spreads in their predictions.

EXAMPLES OF STUDENT THINKING

These interview excerpts are representative of the variety of student thinking that the interview tasks elicited. Although some students got stuck and never moved in their thinking, it was more often the case that as students engaged with these tasks and relaxed a bit with the interviewer they revealed richer levels of their thinking. Sometimes a concept was in transition for a student, and they had difficulty in explicitly articulating their thinking.

Task 1. Brittany (GR7) moves from “don’t know” to “contextual thinking about fruit juice.

BM: Could be like, maybe population or something, well no, because it’s per person.
I: It is per person, very good, that’s a good catch.
BM: Maybe…I know fruit juices are more popular, but what can cause it?
I: Does it surprise you?
BM: I don’t know, kind of, because, kind of puzzling, to me it’s kind of puzzling, it doesn’t really surprise me, it’s just how it happened.
I: So you wonder about it.
BM: Yeah, I wonder about it.
I: Something about fruit juices that would cause it to plunge like that, and then shoot back up?
BM: Shoot back up…um, like, I don’t know, it might be like the trend in how people are drinking, it could be people just started drinking more fruit juice, more choices, more flavors, they got more used to it. Maybe like in 1981 they came out with some really good new flavor.

Task 1. James (GR11) speculates about causes, and tries to relate it to personal experiences.

J: Umm, just what do I notice about it?
I: Yeah. What do you notice?
J: Well it’s steadily going up as the years go by. There are some extreme up and downs but it seems to counter itself out. Every huge jump has a huge down so there’s not that much fluctuation. It seems to have gone down during Desert Storm and up during elections.
I: Up during elections and down during Desert Storm? Which years are you looking at for that?
J: Mainly like 1990 on, because like I’m just trying to recall what I’m drinking and my family.
I: Do you have any ideas why it might have gone down, you said during Desert Storm?
J: Yeah. Umm, I have no idea. Maybe because people weren’t concerned with juice, but it seems like 2000…it seems to go up during election years or around there.

Task 1. Diane (GR12) really has no clue why things are variable in the juice graph.

D: I wonder why it’s going back down.
I: At the end?
D: Yeah. UM…why this is so low [points to the one big dip in 1989] and why this is so high [points to big spike in 1983], but that’s what catches my eye out of the whole thing, is that it hit a high point in 1996, but now its starting to go down again which is really weird considering there’s more people in the United States now and you would think that it would at least stay average if not go up because there’s more people.
I: Now this is consumption per person.
D: Oh yeah uh. [laughs]. Well in that case I don’t know like considering the world nowadays and we’re so Burger King, we want it that way. People don’t sit down and get the right nutrition so I can see it going down, but you’d also think that maybe it would go up just for that reason alone.
I: OK, well you pointed out this spike here what year is that? …. What do you think might be some reasons that might caused the spike?
D: I have no idea. [laughs]. I don’t even know where to begin. That’s really weird, I don’t get either of those (spikes).

Task 1. Elaine (GR12) eventually settles on some reasonable contextual issues to explain variability in juice consumption.

E: People are drinking other beverages, (such as) coffee, alcohol, water.
I: How does that affect fruit juice?
E: They wouldn't be drinking fruit juice as much. I'm not sure when the whole coffee chain started, but maybe if there are more Star Bucks or Coffee People, that people would choose that over fruit juice, so maybe they would drink less fruit juice over the years.
I: Hmm, but you said it was sort of rising right?
E: Yeah [laughs], Umm…. Maybe more juices were made available, more kinds that people liked.
I: What about some of these spikes and dips that you pointed out, what might cause some of those?
E: The markets didn't get the fruit that they wanted and then the next year they did. Oh, or maybe there was something bad in the crops that scared people from drinking juice. I remember hearing about something like that a couple years ago with Odwalla, that juice (company). So maybe it scared people, kind of like mad cow disease now everyone's afraid to eat meat. So then once it past they started drinking juice again.

Task 2. Jack (GR10) chastises the Class A measurers, expects some variability in the data, but not that much, and he is well aware of the potential reliability issue for any future measurements.

JR: It's stupid. How can you measure the same person's foot and get so many numbers? Like it's all spread, someone over here got 17.5 and the most have 22, 23, 24. Outlier, layer whatever you call that.
I: Why do you think they are that different?
JR: I don't know. They just can't measure. I mean, you're supposed to get a variety of answers but when you get something that far off either they didn't measure and guessed or it's just wrong. This is just pathetic (Class A)….
I: Will there be any differences between the classes if they measure everyone's foot?
JR: It's going to be a "big piece of nothing", if this (A) class measured it, there's going to be a whole bunch of off numbers if this class measured it, it's weird. I'd put my money on Class B.

Task 2. Mary (GR7) notices the differences in accuracy in the classes, predicts a wider range would occur in a graph of the entire class's foot measurements, but misses the reliability issue.

MP: Class A is a lot more spread out, they aren't really sure what the correct length is. We had problems with getting a technique like for measuring the head. And it was all over the face, we didn't have a specific place (to measure the head). Or, maybe they were younger….didn't do as good a job at measuring as, say, me (smiles).
(Later) MP: This one seems like they are pretty sure, from 17 to 21 is a really small range, there's no, like, outliers. But between 17 and 26 (Class B) is quite a bit to be off….
I: Suppose that you had like, here is Class A's (entire class measurements) and here is Class B's, what would you expect would be in those (two) graphs?
MP: They'd both be, uh, pretty spread out. They wouldn't be exactly the same, as everyone has different size feet. They'd have the same shape.

Task 2. Ashton (GR11) has no idea why such variations in repeated measurements could occur, and also misses the reliability issue for the entire class measurements *despite* explicit probe attempts by the interviewer.

AT: How did they get from between 17.5 to 25? That's kind of awkward.
I: Why might that have happened.
AT: I don't know how they would get that….
I: (Now asking about measurements for the entire class). What about, what kind of differences would there be between class A and class B? Anything?
AT: Well, are they in the same grade?
I: Let's say yes.

AT: OK, then, um, actually, you know, it's hard to say 'cause people just have, you know, people with different foot sizes, um just have different ones. I'm sure they'd be more similar than if it was a 1st grader and a 12th grader but I'm sure they'd be pretty much in the same range.
I: Does this data [points to graphs] influence the second round of measurements of class wide data at all?
AT: What do you mean?
I: Thinking about what the students did here [points to both graphs in Task 2] and then thinking about their class-wide data, does this data effect how you think the class-wide data would look?
AT: Um, I think so. I mean I think so just because of the, [laughs] it's hard to explain.
I: It is hard to explain. Yea.
AT: Like I could see how, like you could get an idea of what it would look like just because of…(thinks)… well actually no, I don't think so because if this is just one person, and then you're looking at a graph of 25 people or however many people are in the class, you know, it's going to look a lot different, maybe not a lot different but it will look different.

Task 2. Shannon (GR10) starts with observations about individual points, moves on to group comparisons, nails the class-wide measuring reliability issue. Wonders if students were just guessing, or really measuring.

SC: Her foot has to be the size between 23 or 22 because the majority of the class, that's what they got [she circled 23 for class A and circled 22 for class B]…One student got 25 over here [pointed to highest value on graph A] and one student got 24 over here [pointed to highest value on graph B…. It's just weird how, like, a majority of the kids did all spaced out, I mean a majority of the kids, if they weren't lying did [*inaudible*] spaced out.
I: That's on graph A.
SC: Yea, on graph A. Then on graph B, they're all cluttered together. They was all close but they really didn't know. Some people could just be marking what they thought.
I: Would there be differences between the whole class measurements for the two classes?
SC: [Points to graph A] If her foot is just all these different places I don't know what the (whole) class would look like. It would probably look like, mm-mm, like it would look just all scattered around. It would just look like a lot of Xs in a lot of places.
I: Ok, all right what about class B? That was class A.
SC: They would probably be more accurate than class A. 'Cause they look like they know more what they're doing than class A.

Task 3. Keosha (GR10) identifies a range where a majority of the samples should occur, but her reasoning is predominately additive, and she does not explicitly reason proportionally.

I: So you've got some in every category, you think that all of them could happen.
KJ: Umm Hmmm
I: And you've got most of them where?
KJ: Between 71 and 80. Uh, 61 – 70 and 71 – 80.
I: And why *right in that area?* (Note: Here the I: pushes for proportional reasoning….)
KJ: I think that more people would pick from 61 – 70.
I: And so, it's right in 61 – 70. Why 61 – 70?
KJ: Um, '*cause it's a dead possibility for me to pick them*?
I: How do you know?
KJ: I don't know that it…
I: Well I'm saying why 61 – 70 instead of like 51 – 60 or 71 – 80?
KJ: Ah, there's, well, there's more reds, that's why I said that.
I: Yeah but you put 61 – 70, you didn't put like 91 – 100.
KJ: Cause there's going to be at least some yellows, it's not going to be all reds, it could be like 1 red, or 2 reds, there's going to be some yellows.
Task 3. Erica attends to both likely sample proportions, and variation around a center of those proportions, demonstrating distributional reasoning.

I: Why did you put the most in 60 – 70?
ES: Kind of like the central, like, because so many people are drawing them, it would be more likely for a lot of them to get around 60 –70, because the theoretical probability is around 65. A little less than a third, (*wants to check her thinking, reaches for calculator, does computation)*, yeah a little less than a third are in the 50 – 70 range. And then I put about ten students (for) 51 – 60, and 71 – 80, because there is that room for error, well, not error, but some students could get a little less than 65 and some students a little more than 65, and then I put, because there's a bigger chance of getting reds than yellows, I made 81- 90 an 8, and the ones below 51 – 60 less (than 8).

Task 3: Brittany (GR7) appears to be reasoning proportionally, but then talks herself out of it and into reasoning more additively, and drifts off toward 50-50 as the most likely sampling result, ignoring the base rate (population proportion).

I: (Referring to her first 6 samples in part a) Why'd you put those?
BM: Well, I put the 60 and 65, because, well, there's 650 red and 350 yellow, and like if you "took the zeros off" that way it would be more like 100, and there might be some times like you would get like almost the same ratio. And it would vary, so I put like the 60 and the 65. I put the 70 because there probably would be some times when you got like more, and the 35 for the times when most of them are yellow, and not a lot of them are red.
I: (Referring now to the part b) So I noticed that when you were making your guesses (above in part a) you included a 70? What happened down here (points to the table of responses for part b)?
BM: Well, on the 31 – 40 ones, my lowest up here (part a), I put like 5 chances, because once again there's a lot more reds than there are yellows, and, uh, …
I: But you think it (31 – 40) might happen that many times (5)?
BM: Well, yeah, I think so, because *it's almost half and half …*
I: What's almost half and half?
BM: The 31 – 40, well, not 31 – 40, but the 41 – 50, that one's almost half and half.
I: You put the same numbers in there, didn't you (referring to part a).
BM: Yeah, the 41 – 50 and the 51 – 60 one, those two are pretty much half and half if you get one in there, and the 650 and 350, the reds are a little bit over half.

DISCUSSION AND CONCLUSIONS

All three of these tasks were used to assess students' statistical thinking in our interviews, but similar tasks were also used in our classroom teaching episodes to give students opportunities to explore them more deeply. When students have invested time in making their own conjectures about why there is variability in fruit juice consumption over time (Task 1), or why one class's repeated measurements are so different than another classes (Task 2), they are ripe for small group discussion and argumentation about their conjectures. Some of their reasoning will hold up in a group, while other conjectures will be labeled as 'wild speculation.' Those students who have no idea why variability occurs in these tasks will learn from their peers that there are indeed some likely causes for the variation. Predictions for the sampling distribution (Task 3-I), or predictions how many trials are needed until an outlier occurs (Task 3-II), can be tested out by actually carrying out experiments or running computer simulations. In this way we can use tasks that were devised to assess our students' initial statistical thinking about data contexts to create deeper curricular exploration opportunities.

Despite the fact that researchers and curriculum developers have been exploring student thinking and student discourse in order to learn from our students, there has been little attention to student reasoning and justification in most of the statistics courses that we teach in the United States, and this may be the case in other countries as well. As statistics instructors, we are all too quickly caught up in the race to 'cover' material and to get through sections of our textbooks. We do not provide enough opportunities for our students to explore and discuss data, make conjectures, or create their own analysis criteria. Without such opportunities our students will only learn the surface details of statistics, and we will not have clues about their thinking processes. They will not experience the adventure of 'uncovering' the stories that are encrypted in data, and we will fixate on procedures. They will not have a chance to *do* statistics, to engage in their own statistical thinking, and we will not learn where they are in the development of their statistical thinking. It is clear to those of us who have explored students' statistical thinking that students will enthusiastically engage with data sets that are presented in more open ended contexts. There is a gold mine waiting for us in our students' statistical thinking that can provide us with curricular roadmaps to build on in order to further enrich our students' statistical growth whatever their level may be. All we need to do to start the process is to ask, "What do you notice? What do you wonder about?"

ACKNOWLEDGEMETNS

The research project discussed in this paper has been supported by the National Science Foundation (NSF Grant No. RED-0207842). Any conclusions or opinions expressed in this plenary do not necessarily represent the position, policy, or endorsement of NSF.

REFERENCES

Bakkar, A., and Gravemeijer, K. P. E. (2004). Learning to reason about distribution. In J. Garfield and D. Ben Zvi (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*, (pp 147-168). Dordrecht, The Netherlands: Kluwer.

Ciancetta, M. and Noll, J. (2006). Undergraduate students' difficulties assessing empirical sampling distributions. In A. Rossman and B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. Voorburg: The Netherlands: International Statistical Institute.

Mokros, J., and Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education*, 26, 20-39.

Nemirovsky, R. (1996). Mathematical narratives, modeling, and algebra. In N. Bednarz, C. Kieran, and L. Lee (Eds.), *Approaches to algebra: Perspectives for Research and Teaching*, (pp. 197-220). Dordrecht, The Netherlands: Kluwer Academic.

Reading, C., and Shaughnessy, J. M. (2004). Reasoning about variation. In J. Garfield and D. Ben-Zvi (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*, (pp. 201-226). Dordrecht, The Netherlands: Kluwer.

Shaughnessy, J. M., Watson, J. M, Moritz, J. B., and Reading, C. (1999, April). School mathematics students' acknowledgement of statistical variation: There's more to life than centers. Paper presented at the Research Pre-session of the 77th Annual meeting of the National Council of Teachers of Mathematics, San Francisco, CA.

Watson, J. M. and Moritz, J. B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics*, 37,145-168.

Wild, C. J. and Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, 67, 223-265.

United States Department of Agriculture. (2004). *Economic Research Service*. Retrieved www.ers.usda.gov/data/foodconsumption/foodavailspreads.

Zawojewski, J. S. and Shaughnessy, J. M. (2000). Data and chance. In E.A. Silver and P. A. Kenney (Eds.), *Results from the Seventh Mathematics Assessment of the National Assessment of Educational Progress*, (pp. 235-268). Reston, VA: NCTM.