Students' Attention to Variability when Comparing Distributions

J. Michael Shaughnessy, Matt Ciancetta, &Kate Best
Portland State University

Dan Canada, Eastern Washington University

Paper Presented at the Research Presession of the 82nd Annual Meeting of the
National Council of Teachers of Mathematics
Philadelphia, PA
April 20, 2004

Introduction

In the past, K-12 mathematics students' exposure to statistical concepts has been rather impoverished, consisting primarily of measures of center—mean, median, mode—and perhaps some graphical representations of data. Both the *Curriculum and Evaluation Standards* (NCTM, 1989) and the *Principles and Standards for School Mathematics* (NCTM, 2000) advocate for a much wider and deeper role for statistics in school mathematics, including reasoning about data in context and making data-based decisions. An understanding of the role of variability in various contexts—e.g., in sampling, in data sets, and in probability experiments—is one of the critical components students need for statistical reasoning. The research reported here on students' conceptions of variability is part of a three-year NSF grant[1] *The Development of Secondary Students' Conceptions of Variability* (Shaughnessy, 2003).

Variability: A missed opportunity?

Four of the most important ideas in statistics are *centers, samples, comparing data sets*, and *variability*. We pull samples of data to compare the centers and the variability of those samples, and then make inferences about the populations from which those samples were drawn. Until recently, most of the research on students' conceptions of these four ideas was focused on centers. Our research tasks, curriculum materials, and assessment items in statistics seemed heavily weighted toward centers (Mokros & Russell, 1995; Konold & Pollatsek, 2002; Watson & Moritz 2000c), while research tasks and curriculum materials on variability have lagged behind. Batanero et al (1994) also made this point, and noted that Loosen et al (1985) argued that textbooks in statistics put more emphasis on looking at heterogeneity in data than they do looking at variability in data. Green

[1] This research was funded by the National Science Foundation (NSF) under

(1993) also exhorted researchers in data analysis to pay more attention to what students understood about variability. Students' understandings of variability as a concept were not being systematically investigated (Shaughnessy, 1997).

However, in the last few years there has been an increasing interest in research on students' understanding of variability in a number of different contexts. Studies have been made on students' understanding of variability in sampling contexts (Shaughnessy et al, 1999; Reading & Shaughnessy, 2000; Torok & Watson, 2000; Torok, 2000; Shaughnessy et al, 2003) or in probability contexts (Watson & Kelly, in press; Shaughnessy and Ciancetta, 2002). Students' 'definitions' of variability have been investigated (Ciancetta, et al, 2003), as have students' understanding of the variability in sampling distributions (Saldanha & Thompson, 2003). Qualitative analyses, teaching experiments and design experiments have been conducted that document how students handle variability when they encounter it naturally while comparing data sets (Watson & Moritz, 1999; Watson, 2001, 2002; Petrosino, A.J., Lehrer, R., & Schable, L., 2003; Cobb, 1999; Cobb, P., McClain, K., & Gravemeijer, K., 2003). There has even been an attempt to create an assessment tool to measure students' understanding of variability (Watson et al, 2003).

Students are very accustomed to being asked questions in data and chance that prompt them to respond with single point value answers. What is the Probability that….? How many …would you expect? What is the average of…? These types of tasks in both research and assessment have tended to mask the issue of variability. What would students expect to happen if we pulled repeated samples from a mixture or population—what then? Hopefully not the same thing every time, but what sort of distribution or set of outcomes might students expect, and why? What do students attend to when they are asked to make a choice or an inference between data sets, or to describe differences in distributions that reflect variability? This paper presents the results and analysis of two

---

REC-0207842. The opinions expressed in this paper do not necessarily reflect the views

interview tasks that investigated student thinking about variability, one on comparing distributions and another on comparing data sets. These tasks are a small part of a wider exploration of students' understanding of the presence of and role of variability.

Background

The Development of Conceptions of Variability Project is centered at a large urban university in the northwest part of the U.S. Ten mathematics classes in six schools, two middle and four high schools, along with their teachers and students, are the participants in this project. Six of the classes, two middle and four high schools, are research classes, while the other four classes, one middle school and three high schools, are used as comparison classes. Three of the classes, one high school and two middle schools, have been mostly in tact for multiple years of the project, and thus students can be traced over time. Of the six schools, two are urban, three are suburban, and one is rural located over two hundred and fifty miles from the urban center.

The six teachers in the research classes of the project play the dual role of classroom teacher and co-researcher on the project. Similarly, the university research team members are both researchers and co-teachers for the classroom teaching episodes in the project. The six teachers who were invited to participate in the project had been participants in an intensive three-week summer institute on the teaching and learning of probability and statistics funded by the Eisenhower program in the summer of 1999. Thus, the teachers and researchers on this project have worked together on several phases of statistics education over the past five years.

Methodology

Data sources for the Variability project include large-scale class wide surveys, semi-structured task based interviews**,** and designed classroom teaching episodes. Prior to

---

of NSF.

any classroom activity or individual interviews, an initial survey was administered to all students in all ten classes. The tasks on the initial survey were intended to obtain information about student thinking about variability in three principal contexts: in sampling situations, in probability experiments, and in data sets. The initial survey was administered during the first week of class in the fall of 2002. Tasks on the survey included adaptations of tasks previously used by other researchers (Shaughnessy et al, 1999; Reading &Shaughnessy, 2000; Watson & Moritz, 1999; Watson et al, 2003; Watson & Kelly, in press). Reports on student thinking from some pieces of the initial survey have been summarized in Shaughnessy et al (2003 & 2004), and Ciancetta et al (2003).

Following the initial survey, twenty-four students, four from each of the six research classes (two boys and two girls) were randomly chosen for a series of three individual task-based clinical interviews. The first interview focused on sampling and took place at the beginning of the school year in the fall of 2002 after the initial survey, and prior to the first weeklong teaching episode on variability in sampling situations.

The second interview focused on variability issues when comparing distributions or when comparing data sets, and it took place between the first and second teaching episode, in the spring of 2003. Thereafter, the second teaching episode was conducted and involved variability in data sets that the students generated during repeated measurement activities. At the end of the school year in the spring of 2003, a second class-wide survey was given to all the students in all ten classes. The second survey contained many of the items in common with the second clinical interview.

The third interview reviewed issues from the first two interviews, and also included some new tasks that involved variability in graphical representations of data, both univariate and bivariate graphs. The third interview was administered to the students in the spring of 2004, followed by the third teaching episode which focused on variability in large scale data sets of food consumption over time (US Department of Agriculture

data), and data sets about fast food restaurants. Variability in graphical and tabular representations of those two data sets was the major focus of the third teaching episode.

The two tasks that are reported on in this paper, the Comparing Sampling Distributions task, and the Movie Wait Time task, are taken from the second interview, which took place between the first and second teaching episodes.

Subjects and Procedure

Twenty-four students were randomly selected, four from each of the six research classes, to be interviewed. All but one of the 24 students in the second interview had also been interviewed in the first interview. (One student had moved to another state). There were eight middle school students (four from grade 6, four from grade 7) and 16 high school students (four in Algebra 1 (grade 9), eight in Interactive Math III (grade 10 & 11), and four in AP statistics (grade 12)) who were interviewed. The students were interviewed during their regular mathematics class time, each interview lasted about 35 minutes.

The first teaching episode involved sampling tasks from both known and unknown populations. Students were given a scenario that involved randomly selecting a committee of ten members of a school band from a set of 30 boys and 70 girls. One part of the activity asked the students to predict what they thought would occur if 30 repeated samples of size ten were drawn (with replacement) from the 30-70 mixture. What might the distribution of outcomes look like? They graphed their predictions in a stacked dot-plot, and then actually carried out the repeated sampling tasks using containers with colored chips in a 30 – 70 yellow to red ratio, representing males to femles. Graphs of the actual results from the sampling were posted beneath each groups' predicted results for the sampling, and a discussion ensued. Among the issues that often arose in the discussion were students' surprise that extreme values, like 0 or 1 reds, were rare. In another part of the sampling episode, students were given an unknown mixture, an asked

to devise a sampling method of their own by drawing repeated samples, and making a stacked dot-plot of their results which included their prediction of the actual mixture, and their reasoning.

The main point we wish to make is that students had had considerable experience in the first teaching episode in reasoning about, comparing, and making decisions based on graphical representations of the results of repeated samples prior to the two tasks we discuss below from the second interview.

The Sampling Task: Which distributions are real and which are made up?

As mentioned above, the student interviews were conducted after a teaching episode in which students had the opportunity to explore some sampling activities related to the Sampling Problems we discuss here. The students were first presented with the following scenario together with the graphs in Figure 1.

> Working in small groups, students in a class pull samples of 10 candies from a jar that has 1000 candies. They do this 50 times. The jar has 250 yellow and 750 red candies in it. Each time they put the sample back and remix the jar.
>
> Here are some of their actual graphs for the numbers of reds in their 50 handfuls of 10 candies. What do you think about these graphs?

****Insert Figure 1 about here****

This task was given to remind students of the kinds of graphs they had obtained in the classroom teaching episode on repeated sampling during the previous fall, and to see if any of the four graphs would raise questions for them. Then the interviewees were given the following question, along with the graphs in Figure 2.

> Another class did the same experiment, pulling 50 samples of 10 candies from a jar with 750 reds and 250 yellows, and recording the number of reds. However, in this class some of the groups 'cheated' and didn't really do the experiment, they just made up a graph. Here are some of the students' graphs from the class. Which are real, and which are fake? Why do you think this?

****Insert Figure 2 about here****

In the interview setting, students were first asked to offer an initial speculation as to which graphs were real and which graphs were made up and to give reasons for their answers. In the initial responses, fourteen of the twenty-four interviewees identified that two of the four graphs were made up and two were real. The ten interviewees who did not initially identify exactly two as real and two as made up were given the additional information that two of the graphs were indeed real and two were made up. They were then given an opportunity to adjust their initial conjectures and make a final response. The initial number of correct identifications (out of a possible four) for both the middle school and high school student interviewees is displayed in Table 1. Their final decisions are displayed in Table 2.

****Insert Tables 1 and 2 about here****

The results of Table 1 and Table 2 suggest that approximately 41% of the interviewees correctly identified all four graphs in the initial round of responses. After they were given a chance to amend their responses, approximately 51% of all the students were able to identify all four graphs correctly. A finding from Table 2 that warrants mention is that in the final round of responses, middle school students (students in grades 6 & 7) outperformed their high school counterparts (students in grades 9 – 12). Seventy-five percent of middle school participants were able to identify all four graphs correctly while only 43.75% of high school participants were able to correctly identify all four graphs. This may be due in part to the fact that statistics and graphical representations of data were integrated throughout the year in the curriculum of the two middle school research classes. One of the classes used a curriculum (Foreman & Bennett, 1995) that includes significant attention to probability and statistics.

In Table 3 and Table 4, a breakdown of student initial and final identifications for each of the four graphs is displayed (See the Appendix A for a student by student breakdown of responses to each of the four graphs).

****Insert Tables 3 and 4 about here****

These results show that Graph B was the most commonly misidentified by both middle school and high school students. Ten out of 24 students misidentified Graph B in their final response as a real graph. Both in their initial responses and their final responses, nearly all students identified Graphs D and C correctly. Students' reasoning and explanations for their choices is addressed in the next section of this paper.

Student Reasoning about the Sampling Task

As students were reasoning about which distributions were real and which were made up in Figure 2, their reasoning predominantly fell into four types: paying *attention to the extremes, attention to shape, attention to spreads,* and *attention to centers,* roughly in that order of frequency.

Attention to Extremes

The most common type of argument used to support students' final responses regarding which graphs were real and which graphs were made up involved their attention to extreme values in the distribution. In particular, attention to extreme values was used by almost all students in determining whether or not Graph A, B, and/or C were real. Eight students pointed out that there were a lot of handfuls of ten reds in Graph A. Five of the same students also pointed out that there were no handfuls of ten reds in Graph B. In the following passage, J.L. (a 11th grade student) describes his reasoning regarding the number of ten reds handfuls in Graph A and Graph B and why Graph A should be fake and Graph B should be real.

> J.L.: B seems real to me because…there are no tens, which I, because looking back at like when you were in our class, we never got one sample that was all one kind…and A, it just seems like there are way too many tens for it be real. Yeah, I don't think you'd have six consecutive, or not consecutive, but six times where it's all going to be red.

J.L. finds it reasonable that Graph B would have no handfuls of ten reds because he views a handful of ten reds as somewhat extreme. J. L.'s comments about the number of handfuls of ten reds in Graph B are revealing when they are juxtaposed with his thoughts regarding the number of handfuls of five reds in Graph C. In the following passage J.L. considers the absence of handfuls with five reds in Graph C:

> J.L.: C, it doesn't seem that it would be made up to me because they left out five, and I just, I wouldn't do that because it seems likely to me that I would get at least one five out of the mixture which is kind of a 50-50 thing. But, ah, that never happens so it would be more genuine than D…It seems to me usually that no matter what I do with these types of samples that there's going to be at least one time where I'll get half reds and half yellows, or half whatever and whatever, and that would be five and five…It seems that if you were going to make up numbers you would throw at least one (handful of five reds) in there.

J.L. comments that, on the one hand, in Graph B no handfuls of ten reds seems reasonable to him. On the other hand, in Graph C the absence of handfuls of five reds is suspicious. Earlier in the interview, J.L. had the opportunity to describe what a theoretical distribution might look like:

> J.L.: If I were going to cheat I would do a more progressive thing, like where would it be in its peak in the proportion and then kind of filter down from there…because the average, since it's not a 700 or 800, it's right in the middle, in my mind, like if everything came out perfect these two (the number of handfuls with seven reds and the number of handfuls with eight reds) would probably be exactly the same…and then it just slowly comes down. (student uses fingers here to trace down both sides from the mode).

Following this line of reasoning, J.L. might conclude that handfuls of ten reds and five reds would have about the same frequency, but this did not occur when he considered Graph B and Graph C.

Nine other students mentioned that in Graph C, there was no trial that resulted in a handful with five reds. Three students, including J.L. claimed that this was troubling because five was an 'average number' and that you should get at least one five in fifty trials. When asked what she meant by 'average' in this case, D.P. (an 11th grade student) clarified, "it's the average number between zero and ten". For the three students for whom this was a concern, the actual composition of the mixture (750 reds and 250

yellows) seemed to have little bearing on the idea that one should be able to pull a five red / five yellow handful. This is reminiscent of many other studies in which students were found to ignore base rates when making decisions under uncertainty (e.g., Tversky & Kahneman, 1974).

The number of handfuls with only three or four red candies in Graph B was also a concern for many students. Fourteen students attended to the large number of threes and fours in Graph B. Attention to the extreme values of Graph B, however, was not enough to convince six of the fourteen students that Graph B was made up. This was because those six students believed that either the absence of fives in Graph C or the abundance of tens in Graph A were indications that Graph C or A was less likely to be real than Graph B. However, after the students were told that exactly two of the graphs were made up, 12 of 14 students determined that Graph B must be one of the fake ones.

Attention to Shapes of the Graphs

Of the four graphs, Graph D appeared to be the easiest for most students to discern that it was one of the two made up graphs. Almost every student (23 out of 24) correctly identified Graph D as being made up. Furthermore, most students attended to Graph D first in their interview responses. There were 17 out of 24 students interviewed who attended to Graph D first and said things like Graph D "looked the fakest", "was definitely made up", "was cheesy fake", "was the least real looking", or "was too perfect". When asked to clarify why they believed Graph D to be made up, students cited two major factors in their decision making. Fourteen of the students cited the shape of Graph D as problematic, and 11 of the students cited that Graph D was too spread out or had too many low numbers (like twos and threes). Regarding the shape of the curve specifically, many students claimed that graph D was a "perfect Bell curve", or was "too symmetric", or "had too much of a pattern", or "looked like a staircase", or "looked too

accurate". In the following passage, M.P. (a 7th grade student) explains the pattern that she saw when she looked at Graph D:

> M.P.: D looks made up because, I know when I'm making graphs you sort of start with a really small one and then you get bigger – not necessarily the same amount. But, like this one, it goes up by one, up by one, then up by two, then up by two, then up four, then it stays, and then it goes down five, and then down by two.

M.P. was the student who said Graph D was "cheesy fake."

Although many students referred the Bell curve shape of Graph D, comments regarding the Bell curve were also cited in reasoning about Graphs A and B too a lesser extent. The Bell curve explanation was used by five students to support why Graph B or Graph A might be real. These five students liked the fact that Graph B and/or Graph A had a 'general Bell curve', and that this would be something that they would expect from a real graph. Thus, some students expected the data to be normally distributed but were simultaneously bothered by a 'perfect' looking distribution.

Occasionally, a student's attention to the shape of the distribution would compel him or her to offer an incorrect response to whether a graph was made up or real. For instance, S.C. (Grade 9) appropriately determined that Graph D was made up because "they marked going in a pattern", but when she looked at Graph A she said, "Graph A, it looks like they just put three more or two more, three more x's each time, or three less x's". When she was asked to clarify why she liked Graphs B and C better than Graph A (as the real ones), she mentioned that in Graphs B and C "the majority of the reds is seven, eight, and nine…and the majority of this one (reference to Graph A) is like nine and seven…on Graph B it was six, seven, and eight." Although S.C. is using the term 'majority' here, she seems to be referring to the modes or clusters of peaks in each graph. She is bothered that in Graph A the mode is at seven red candies but there is a dip in the frequency of eight red candies before the graph spikes again at nine red candies.

Attention to Spreads of Graphs

As mentioned in the above section, 11 of the 24 interviewees cited that Graph D was too spread out in the distribution. Those students who focused on the exaggerated spread in Graph D attended mostly to the frequency of the extreme values, like the fact that there is an incident where one student pulled out only two reds in a handful of ten candies in the distribution of Graph D.

Spread was also a deciding factor for many students in Graph A, B, and C. Thirteen students cited an appropriate spread in which they would think that most of the number of reds in each handful would occur. The 13 students mentioned either that most of the samples should fall between six to nine red candies in a handful. Most students tended to have more difficulty negotiating which of graphs A, B, and C was made up and which was real. Eight students used 'an appropriate spread' argument to support that Graph A was real, three students to support that Graph B was real, and eight students to support that Graph C was real.

Attention to Centers of the Graphs

A tendency to consider the centers of the graph or to demonstrate proportional reasoning did not explicitly emerge as frequently in the interview as the other the other three types of reasoning discussed above. Only eight of the 24 students explicitly attended to the modes of at least one of the four graphs. In five of these cases, students considered the mode of Graph B when determining whether or not Graph B would be a real graph. In the following passage, C.W. (Grade 12), discuss the mode of Graph B being seven, and the implications of that for his decision to rate Graph B as either real or fake.

> C.W.: I think that possibly Graph B might be one of the made up ones just because in all of these ones, seven is never the highest. And so you know, if I was looking at something I would always say, "You know, well seven hundred and fifty – that's about 7.5% or 75%" so I'd say it would be around seven or eight and they kind of spiked them both – seven and eight were kind of the highest. Now it could happen, but I mean, I'm thinking from the perspective of trying to make it look like it's real. (It looks) too normal, maybe too perfect for a sample of ten…Because the groups cheated, they kind of want to make it to look normal, but not like seven and eight are the

highest.

A little later in the interview C.W. finalized his choice of Graph B as made up. He argued that while the number of six red, seven red, and eight red handfuls should be the highest, you would also make them the highest if you were trying to make your graph look real when it really wasn't. His choice, to include Graph B as fake, hinged on the idea of the center of B looking 'too real'. In his decision-making, C.W. never attended to the extreme values in Graph B to support his decision. It is also important to note that although C.W. would have been counted in Table 3 and Table 4 as having the correct response on Graph B as made up, his reasoning lacked explicit attention to the extreme values in Graph B. On rare occasions, students such as C.W. would offer a correct response to whether or not a graph was real, but not be able to support their decision with convincing evidence that they really understood the roles of various perspectives like shape or spread in this problem

Conclusions about the Sampling Task

In general, students performed very well on the Sampling Task. Many students were able to offer all four correct responses when interpreting which graphs were real and which graphs were made up. Some students only responded correctly regarding two graphs but no student responded (either initially or finally) with every graph misidentified. Despite the fact that occasionally there was a lack of appropriate reasoning to accompany a response, many students attended to appropriate aspects of the graphs when making their decisions.

Overall, students had a tendency to rely mainly on extreme values when determining whether or not a graph was real or made up. This was particularly true for Graphs A, B, and C in which issues arose regarding how many threes, fours, fives and tens one could expect to see in a distribution. As mentioned earlier, the students had the opportunity to explore a teaching episode in which they performed similar sampling tasks

prior to the interviews. In this episode, many students held the misperception that extreme values weren't necessarily 'extreme'. Students tend to believe that extreme values will occur a lot more often than they actually will. After students performed sampling activities in class and ran simulations, many students were surprised at how rarely samples with extreme values occurred. This may have made a significant impact on students and it might offer a partial explanation as to why so many of these students paid attention to extreme values in this task. Many students used very appropriate and well-articulated reasoning regarding extreme values in this task, and some directly referenced what they witnessed in the teaching episode as influencing their decisions about the graphs.

In Graph D, students tended to focus more on the spread of the distribution or the shape of the distribution. Students did not often attend to centers and with a few exceptions, when they did refer to centers, it didn't impact their decision making significantly. Based on the similarity of the centers among the four graphs, we were not surprised by the students' inattention to centers.

The Movie Wait Time Task

The results from the task reported here involve students' reasoning about two data sets in a context and their data-based decisions. The task introduces the notion of *movie wait-time* as the difference between the advertised start time of a movie (like in the news paper) and the movie's actual start time. Two data sets, along with their corresponding stacked dot plots, display wait-time data for two movie theater chains. The data sets had identical means and medians of 10 minutes (this information was printed below each data set) and both were bi-modal with different spreads. In particular, the data set for

Maximum Movie Theaters had a range that was wider than the data set for *Royal Movie Theaters*. The task is shown in Figure 3.

****Insert Figure 3 about here****

The students were given questions in three parts and asked to reason and make decisions about the movie wait-times for two theater chains.

> Part A: "What can you conclude about the wait-times for the two theaters?"
> Part B: "One student in the class argues that there is really no difference in wait-times for movies in both theaters, since the averages are the same. Do you agree or disagree? Why?"
> Part C: "Which of these theaters would you choose to see a movie in? Why?"

Each interview was video taped and transcribed. Transcription segments of each student's responses fell into six categories: *Specific Data Points, Variation, Center, Distribution, Informal Inference,* and *Context*. The transcription segments were categorized according to the descriptive language used and/or reference to calculations and other facts or observations. Characterizations of these categories along with some example segments of student responses follow. The part of the task that the student is responding to is noted above each example.

Specific Data Points: The student compares or examines specific data points, such as listing off data points or comparing individual data points or wait-times. The student may specifically compare the high ends of the distributions, or the low ends, or both.

> <u>Part A</u>
> KJ (grade 9): [*points to Maximum data set*] this one has 14, 13, 12, 11 and so on, and this one [*points to Royal data set*] only has 11, the closest one is 11.5 and that's, I think that's the highest.

> <u>Part A</u>
> BC (grade 7): [*points to ROYAL data set*] there is a lot more 8's, 9's and 10's and 11's than there is [*points to Maximum data set*]. Well, this one [*points to Maximum data set*] has some 14's and 13's and some of the high numbers but there's a lot of low numbers also.

Variation: The student refers to clusters, relative amounts of variation or spread, the exact or approximate ranges of 9 for Maximum and 3 for Royal, or the exact spreads of 5 to 14 for Maximum and 8.5 to 11.5 for Royal.

> <u>Part A</u>

SA (grade 7): Well, [*points to Royal graph*] it's within an 8.5 and 11.5 range. Up here [*points to MAXIMUM graph*] you've got some all the way down to 5 and all the way up to 14.

Part A
JM (grade 11): Well, their [*Royal*] wait-times are kind of bunched up together and, and Maximum movie theaters is more spread out.

Center: The student refers to the means or medians or both (Note: No students referred to modes, perhaps because the distributions were purposely made bi-modal. The purpose behind having the students reason about bi-modal graphs was to prevent them from making decisions based on a comparison of modes.)

Part A
TP (grade 6): Well, both of them they have a mean and median of about 10 minutes of wait-time.

Part B
JL (grade 11): I agree because, well the averages are the same.

Distribution: The student refers to both the centers and variation simultaneously.

Part B
MH (grade 11): at the Maximum movie theaters it could be just 5 minutes and not 10 minutes, which is half the average wait-time.

Part B
KL (grade 12): Um, you also have to like look at how much it's weighted because there's, because this one [*points to high end of Maximum graph*] is so much higher and that would account for the smaller ones, like that would bring down the average versus these ones [*points to Royal graph*] there is just more of them, like these ones [*points to Maximum graph*] were more spread out and it kind of evened out.

Part B
KH (grade 12): I would, I would disagree because when you can see the graphs [*points to Maximum graph*] even though the means and medians are the same, when you look at the graphs it looks very different, so you'd have to look for standard deviation how like it is because if you went to Royal movie theaters, you would be more likely to watch a preview that's exactly at 10 or very close to 10 than you would be at Maximum movie theaters because not many are close to 10 you would either be a lot more likely to watch a preview that was shorter than 10minutes and then average it with your friend's who is higher to get the time but not exactly 10.
I: So how would you, since you mentioned standard deviation, how would you use that?
KH: Um, well you just, to find on average how far away each X is from 10 and so since that the most any movie time for Royal movie theaters is away is 2, they're all under, within 2 minutes of 10 minutes and then, um, these ones there within 5 minutes higher or lower at Maximum movie theaters, so the standard deviation would be higher because even though you would expect to, on average find one for 10, it's unlikely that just any one trip would be at 10 minutes.

Informal Inference: The student makes inferences about the probability of experiencing certain wait-times, or makes statements about expectations or uses language that refers to predictability, consistency, reliability or chance and luck or lack thereof.

Part B

JP (grade 11): So if you ask someone what's the average, you tell them it's going to be the same so then what you'd expect is going to be 10 minutes and that would be, you could go up or down from there, but that would be what you'd expect.

Part C
EG (grade 9): lowest is 5 minutes [*for Maximum*], I could get lucky and get 5 minutes or I could not be very lucky and get 14...I'd rather have 5 1/2 than 8 1/2 if I ever got lucky

Part C
HV (grade 6): I would probably choose Royal movie theaters because, um, I would be, I would know how long the wait-time would be, I would know what to expect in wait-time and with Maximum movie theaters I, I wouldn't know if, ah, it would be, if I liked to have really short wait-times, I wouldn't know how long it would be, I wouldn't if it was going to be 5 minutes or I wouldn't know it was going to be around like around 15 minutes of wait time, a, it, you have, you don't know what to expect.

Context: Student speculates about reasons for differences or variation in wait-times. Student mentions his or her own personal preferences or opinions about wait-times or personal experiences with wait-times.

Part C
DP (grade 11) - the longest wait that they [*Royal*] have is 11 1/2 and I'm willing to wait that long. With 14 minutes I'd kind of argue like, "OK let's get on now" *(also coded under Specific Data Points – comparing ends of distribution)*.

Part C
CW (grade 12) - normally I don't want to go to a bad movie and so I'm thinking, you know, if it's a bad movie they'd probably play less commercials so the good movies, you know, at the ROYAL theater would probably be, would have the most commercials and the most commercials are only 11 1/2 minutes but the most commercials at the MAXIMUM, at the MAXIMUM theaters are 14 minutes. *(Also coded under Specific Data Points – comparing ends of distribution)*.

Results for Part A: "What can you conclude about the wait-times for the two theaters?"

The intent behind asking this open question was to find out what the students attended to when left to their own devices.

****Insert Table 5 about here****

Any one student's response could have segments that fall into multiple categories. For example, in responding to part C, Tim, a sixth grader, chose Royal. A particular segment of his response fell under *Informal Inference* and *Variation*. The segment is as follows:

T: Probably Royal.
I: Why?
T: Because it's more predictable. It's between eight and eleven and a half.
I: OK
T: So, you'll probably know it's about ten minutes.

Tim made the *informal inferences* that Royal's times were "more predictable" and "about 10 minutes", he also referred to *variation* by approximating the spread for Royal's data at "between eight and eleven and a half". Most students responded in multiple ways, just as Tim did. So, adding up the students in each row will not add up to the total number interviewed (eight middle school students,16 high school students and 24 students overall).

As seen in Table 5, a sizeable majority of these students' responses (70.8%) were categorized under *Variation*. The second most common category was *Centers* (58.3%). The information that both the means and medians were 10 accompanied the data and also was frequently pointed out by the interviewer. Thus for the ten students who did not explicitly mention that the centers were equal, the underlying assumption was that they at least were aware of it. Every student either made mention of the variation in the graphs or mentioned that the centers were equal and seven students gave responses that overlapped in both the *Variation* and *Centers* categories (two MS students and five HS students). Five of those seven students talked about variation and centers separately, in a compartmentalized way, and so those responses were not classified under *Distribution*. In the *Specific Data Points* category all but one of the eight students either examined an end point of a data set or compared the ends of the distributions. Except for one student, all who made some initial *Informal Inferences* either noted the predictability or consistence or reliability of the Royal movie wait-times, or they mentioned that the Maximum wait-times were unpredictable/sporadic/flexible. That one student referred to the Maximum movie wait-times and said, "when all is said and done you have to wait about 10 minutes". For this student, it didn't matter which theater you went to, you'd have to wait ten minutes. Finally, of the eight students that addressed *Context*, four gave reasons for the differences in wait-times, two talked about their own experiences wait-times and five gave their own preferences such as which theater they would go to.

Results for Part B: "One student in the class argues that there is really no difference in wait-times for movies in both theaters, since the averages are the same. Do you agree or disagree? Why?"

The methodology of this question follows that used by Watson (2002) where students' are asked to defend or oppose an argument that may be commonly heard in a classroom situation.

****Insert Table 6 about here****

No middle school students and only three high school students agreed there was no difference in the wait-times for the two theaters. The decisions that those three students made seemed to rely on the fact that the centers for the two data sets are equal. One middle school student and two high school students would not commit to agreeing or disagreeing, thus were coded *amBivalent* (B).

The majority of both middle and high school students argued that there really are differences in the wait-times for the theaters. Twelve of the 18 students that disagreed cited *variation* in the data sets. These students tended to either state comparisons of the specific spreads for Maximum (5 – 14) and for Royal (8.5 – 11.5) or focus their arguments entirely on Maximum's data having more variation, i.e. "more spread out". Of the remaining six students' who disagreed, all made *informal inferences* and five of those six students' responses also overlapped into the *variation* category

Results for Part C: "Which of these theaters would you choose to see a movie in? Why?" This part of the task personalized the context for the students. Responses are summarized in Table 7.

****Insert Table 7 about Here****

None of the students specifically referred to the centers of the distributions, although students making informal inferences might have implicitly considered centers in their thought process. Given a choice of theaters to attend, just over 20% chose Maximum, just over 70% chose Royal and just under 10% were ambivalent.

Of those who chose to go to the Maximum movie theater, most provided reasons that fell under *Informal Inference*. Those students could be thought of as the *risk takers* as they specifically reasoned that there is a chance of getting a very low wait-time, i.e. five minutes, at the Maximum theaters. They were willing to risk waiting a long time (14 minutes) for the chance to wait a short time (5 minutes).

Only two students did not commit to a decision and thus were coded ambivalent. The discussion with one of those students continually drifted off task and he never made a conscious choice. The other student was adamant that it didn't matter which theater he attended. This student's reasoning was very consistent as he also choose neither agree nor disagree (coded ambivalent) with the statement that there is no difference in wait-times for the theaters. He did notice the variation in the data sets but considered it non-influential in his decision process.

Finally, of the students who chose to go to the Royal theaters, a large majority, 14, of these students used *informal inference* in reasoning about their choice and could be thought of as the *planners*. The *planners* desire to know exactly when the movie is going to start so that they can plan other activities or plan to miss the commercials, i.e., [*at Maximum*] "if you have to go to the bathroom before the movie started, you wouldn't know when it was going to start, you wouldn't know how much time you had or if you wanted to get food or something". Their arguments primarily focused on predictability of the wait-times for the Royal theaters, i.e. [*at Royal*] "I have a really good idea of what time I could wait. I could wait the 10 minutes longer and I wouldn't miss hardly any of the movie or I'd be there for a minute and 30 seconds seeing previews or trailers". Five of the six students who supported their choice for Royal with reasoning about *variation* also

were among the 14 who made *informal inferences*. The various *variation* arguments indicated that those students saw that the Maximum data displayed more variation than the Royal data.

Interactions

As one may suspect, by the large number of students who disagreed that the wait-times were the same in part B and chose to go to the Royal theaters in part C, the most common pairing of decisions were Disagree-Royal at 58.3%. Four students (16.7%) chose the Disagree-Maximum pairing, two students were ambivalent on part B but then chose to go to Royal and one student each chose the pairings Agree-Maximum, Agree-Royal, Agree-amBivalent, and amBivalent-amBivalent respectively. Overall 14 of the 24 students provided reasoning about *variation* and also made some *informal inferences* regardless of the decisions they made on parts B and C.

The Disagree-Royal (D-R) pairing was the only pairing for which the students' reasoning combined from both parts B and C fell into each of the six categories (This was due to overlaps.) Ten of the 14 students who responded D-R gave combined reasons that addressed both *variation* and made *informal inferences*. Those ten could be considered the *hard-core planners*. They saw differences in the data sets and chose to go to the Royal theater. The *hard-core planners* all inferred either that Royal's wait-times were more the more predictable of the two, or that Maximum's wait-times were unpredictable or both. They seemed to attend to the tighter spread in Royal's data set.

All four of the students who made the Disagree-Maximum (D-M) choice also provided reasoning that fell in the *Informal Inference* category. Each of those students argued that at the Maximum theater, there is a chance of getting a wait-time that is lower than the lowest Royal wait-time. Two of the students who chose D-M also gave combined reasons that addressed both *variation* and made *informal inferences*. These two students could be considered the *hard-core risk takers*. Similar to the *hard-core plannars*,

the *hard-core risk takers* also saw differences in the data sets, such as Royal's tighter spread and Maximum's wider spread, but they choose to go to the Maximum theater. They seemed to focus on gambling on the possibility of experiencing a very short wait-time at the Maximum movie theater.

Results and Discussion of the Movie Wait Time Task

An examination of responses to all the three parts of this task revealed:

- 15 students provided responses that were categorized under *Specific Data Points*. Thirteen of those students gave responses that indicated they compared either the high end points of the distributions or the low end points of the distributions or both.
- 22 students gave responses falling under the *Variation* category. Seventeen of those 22 specifically mentioned the spreads or ranges of the data sets.
- 16 students referred to the equal means and medians of the data sets with none of the responses coming from part C. Certainly, these students did not consider *centers* when deciding which theater they would attend.
- Only six students addressed the *distributions* of the data sets. Two of the four 12th graders gave the most sophisticated responses about the distributions. The others responses were generally naive.
- 22 students made at least one *informal inference* during the interview. Only one of these 22 students did not make an inference when responding in part C.
- Just over half, 13, of the students' responses fell into the category of *Context*. These students either mentioned reasons for why wait-times would be different, their own preference for shorter wait-times or apathy about wait-time. Some students also discussed their own wait-time experiences.

The categorizations of responses to part A indicate that the students initially attended to both centers and variability in the data. That most students did not attend solely to the centers to the exclusion of variation is an encouraging result in this section. It's also interesting that one-third of the students discussed the context of the data on their own with out any prompting. An unanswered question that arises from these observations is: Did the context of movie wait-time influence students to reflecton their own experiences and preferences, which in turn influenced them to attend to the variability of the weight times in the data sets?

The responses in part B indicate that three-quarters of these students see the data sets as different even though they had the same mean and median. The students who felt the data sets were not the same defended their decision with reasoning that fell in to every category. Ten of these 17 students provided reasoning that was categorized in multiple ways. The majority of categorizations were linked to *Variation*. Noticing and quantifying (in terms of "more" or "less") seemed to play a key role in these students determination that the data sets were different. Also, the use of language aided in the categorization of responses, thus an alternate or even finer grained analysis of responses may indicate that even more students supported their decision with reasoning related to variation.

Of the five students who chose to go to the Maximum theaters, four had previously indicated that they also considered the wait-times at the two theaters to be different (the D-M students). Each of these four students made *informal inferences* concerning the possibility of waiting a very short time at the Maximum theater. Two of the D-M students also relied on reasoning that fell under *Variation* to support their decision to disagree. These two students are the *hard-core risk takers* as they seem to view the data sets as different but consciously decide to go to the Maximum theater for the chance of waiting a short time at the risk of waiting a long time.

Just over 70% of these students chose to go to the Royal movie theater. Their reasons generally indicated a higher confidence in experiencing a 10-minute wait-time at

the Royal theater. Additionally, 14 of these 18 students who chose Royal previously had indicated that they did consider the wait-times at the two theaters as being different (the D-R students). Ten of the D-R students specifically reasoned under both *Variation* and *Informal Inferences*. All the *informal inferences* that these 10 students made showed some indication that their inferences are in part based on their interpretation of the variation in the data sets, these students are the *hard-core planners*. For example, one seventh grade student, Mary, chose Disagree in part B mostly based on a comparison of the spreads of the two data sets, she then chose Royal in part C indicating that she could predict about a 10 minute wait at the Royal theater but couldn't be sure at Maximum. The reasoning provided by the *hard-core planners* suggests a relationship between reasons coded as *Variation* and reasons coded as *Informal Inference* , which may warrant further investigation.

Final Reflections

The research project reported on here is focusing on student reasoning about variability. Do students attend to variability in data sets, and if so how? Do students appeal to variability when making inferences about distributions, or in comparing distributions, and if so, how? The results from these two tasks, Comparing Distributions and Movie Wait Times, provide ample evidence that students do indeed attend to variability, especially when the role of centers is not very salient in the distributions or data sets that they are comparing. All of the distributions in the four graphs in the first task (Which are real and which are made up?) had very similar centers. Certainly if students were making up a distribution of outcomes from a 75% red population, one would expect mean-median-mode to be nearby that percentage in the made up distribution as well. In the movie wait time task, the mean and median were identical, and so centers were not a salient issue in the comparison of the data sets. The research goal

was to shine the light on variability in these tasks, to give students an opportunity to reason using variability when comparing data sets.

A number of conceptions about variability arose in the student thinking on these two tasks. Variability as extremes or possible outliers; variability as spread; variability in the heights of the columns in the stacked dot plots; variability in the shape of the dispersion around center; and to a lesser extent, variability as distance or difference from expectation, all these conceptions of variability surfaced in the students' responses to these two tasks. These are indicators that students can attend to and reason about variability, if they are given a chance to do so. Statisticians such as Moore (1997) and Wild & Pfannkuch (1999) have suggested that variation is the lynchpin of statistics, and without it, there is nothing to investigate. Their writing suggests that variability should be the issue around which we organize the teaching of statistics. The research presented here indicates that aside from the statistical reasons for emphasizing variability in teaching statistics, there may be some sound pedagogical reasons as well. Students have some intuitions about variability that might provide a foundation to build upon. This seems particularly true when the data sets they are considering are within a familiar context such as the Movie Wait Time data.

For future research, it would be interesting to see if the categories under which student reasoning fell on these two tasks will generalize to other tasks and other contexts. Of particular interest is whether students reason about distributions and data sets by integrating both centers and spreads simultaneously, or whether their thinking is compartmentalized, and arguments using centers or variability occur separately without integration on a task. This points to the tension that occurs between expectation and variation when we compare data sets or distributions. Which should take priority, and in what types of situations does one or the other play the larger role? When are both necessary to make a statistical decision? Watson & Kelly (in press) discuss the tension between expectation and variation when students are reasoning about probability

experiments. Shaughnessy et al (in press) discuss additive, proportional, and distributional thinking when students are reasoning about repeated sampling situations, such as the Comparing Samples task in this paper. In distributional thinking, the issue of appropriate dispersion or spread around centers arises as students attempt to integrate both concepts simultaneously.

Based on this study, and on the work of others referenced in this paper, we feel that the area of students' conceptions of variability should continue to provide fertile ground for research into the learning and teaching of statistics for some time. The work has just begun.

References

Canada, D. (2004). Pre-serivce elementary teachers conceptions of variability. Unpublished doctoral dissertation. Portland, OR: Portland State University.

Ciancetta, M., Shaughnessy, J.M., & Canada, D.(2003). Middle school students emerging definitions of variability. In N. Pateman, B. Dougherty, & J. Zilliox (Eds.). Poster Session in the *Proceedings of the 27th Conference of the International Group for the Psychology of Mathematics Education: Vol. 4* (p. 481). Honolulu, HI: University of Hawaii.

Cobb, P. A. (1999). Individual and Collective mathematics development: The case of statistical data analysis. *Mathematical Thinking and Learning, 1,* 5 – 44.

Cobb, P., McClain, K., & Gravemeijer, K. (2003). Learning about statistical co-variation. Cognition and Instruction, 21(1), 1–78.

Foreman, L.C, & Bennett, A. B. (1995). *Visual Mathematics*. Salem, OR: The Math Learning Center.

Konold, C, & Pollatsek, A. (2002) Data analysis as a search for signals in noisy processes. *Journal for Research in Mathematics Education*, *33,* 259 – 289.

Loosen, F., Lioen, M., & Lacante, M. (1985). The standard deviation: some drawbacks of an intuitive approach. *Teaching Statistics, 7*, 29-39.

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

Moore, D. (1997). New pedagogy and new content: The case of statistics. *International Statistics Review, 65,* 123 – 165.

National Council of Teachers of Mathematics, (2000). *Principles and Standards for School Mathematics*. NCTM: Author

National Council of Teachers of Mathematics, (1989). *Curriculum and Evaluation Standards for K-12 Mathematics*. NCTM: Author

Petrosino, A.J., Lehrer, R, & Schable, L (2003). Structuring error and experimental variation as distribution in 4th grade. *Mathematics Thinking and Learning, 5,* .

Reading, C. & Shaughnessy, J. M. (2000). Student perceptions of variation in a sampling situation. In T. Nakahara & M. Koyama (Eds.), *Proceedings of the 24th Conference of the International Group for the Psychology of Mathematics Education: Vol. 4* (pp. 89 – 96). Hiroshima, Japan: Hiroshima University.

Saldahna, L. & Thompson, P. (2003). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51,* 257 – 270.

Shaughnessy, J. M . (2003).*The Development of Secondary Student's Conceptions of Variability*. Annual report year 1, NSF Grant No. REC 0207842. Portland, Oregon: Portland State University.

Shaughnessy, J. M., Ciancetta, M., & Canada, D. (in press). Types of student reasoning on sampling tasks. To appear in *The Proceedings of the 27th meeting of the International Group for Psychology and Mathematics Education*. Bergen: Norway..

Shaughnessy, J. M., Ciancetta, M., & Canada, D. (2003). Middle school students' thinking about variability in repeated trials: A cross -task comparison. In N. Pateman, B. Dougherty, & J. Zillah (Eds.). *Proceedings of the 27th Conference of the International Group for the Psychology of Mathematics Education: Vol. 4* (pp.). Honolulu, HI: University of Hawaii.

Shaughnessy, J. M., & Ciancetta, M. (2002). Students' understanding of variability in a probability environment. In B. Phillips (Ed.). *CD of the Proceedings of the Sixth International Conference on Teaching Statistics: Developing a statistically literate society, Cape Town, South Africa*. Voorburg, The Netherlands: International Statistics Institute.

Shaughnessy, J. M., Watson, J. M, Moritz, J. B., & 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.

Torok, R. (2000). Putting the variation into chance and data. *Australian Mathematics Teacher, 56,* 25 – 31.

Torok, R. & Watson, J. (2000). Development of the concept of statistical variation: An exploratory study. *Mathematics Education Research Journal, 9,* 60 – 82.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124-1131.

Watson, J. M. (2002). Inferential reasoning and the influence of cognitive conflict. *Educational Studies in Mathematics, 51, 225 – 256*.

Watson, J. M. (2001). Longitudinal development of inferential reasoning by school students. *Educational Studies in Mathematics, 47,* 337 – 372.

Watson, J. M., & Kelly, B. A.(in press). Expectation versus variation: Students' decision making in a chance environment. *Canadian Journal of Science, Mathematics, and Technology Education*.

Watson, J. M, Kelly, B.A., Callingham, R.A., & Shaughnessy, J. M. (2003). The measurement of school students' understanding of statistical variation. *International Journal of Mathematical Education in Science and Technology, 34,* 1 - 29.

Watson, J. M., & Moritz, J. B. (2000). The longitudinal development of understanding of average. *Mathematical Thinking and Learning, 2,* 11 – 50.

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

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

Appendix A

Individual Students' Initial / Final Responses on Whether the Graphs *Were Real or Made Up*

	Graph			
Student (Grade)	**A**	**B**	**C**	**D**
KH (12)	√ / √	√ / √	√ / √	√ / √
BN (12)	**0 / 0**	**0 / 0**	√ / √	√ / √
KL (12)	√ / √	√ / √	√ / **0**	√ / **0**
CW (12)	√ / √	√ / √	√ / √	√ / √
JP (11)	√ / √	√ / √	√ / √	√ / √
EL (11)	√ / √	√ / √	√ / √	√ / √
MH (11)	√ / √	√ / √	√ / √	√ / √
MD (11)	**X / 0**	**0 / 0**	**X** / √	√ / √
JM (11)	√ / √	√ / √	√ / √	√ / √
MF (11)	√ / √	**0 / 0**	√ / **0**	√ / √
DP (11)	**0 / 0**	**0 / 0**	√ / √	√ / √
JL (11)	**0 / 0**	**0 / 0**	√ / √	√ / √
JQ (9)	√ / **0**	√ / **0**	**0** / √	√ / √
KJ (9)	**0 / 0**	√ / **0**	√ / √	√ / √
EG (9)	**0** / √	**0** / √	√ / √	√ / √
SC (9)	**0 / 0**	**0 / 0**	√ / √	√ / √

Note. √ refers to a correct response, 0 refers to an incorrect response, and X refers to when there was no data for a students' response

	Graph

Student (Grade)	A	B	C	D
MP (7)	√ / √	0 / √	√ / √	√ / √
AG (7)	√ / √	√ / √	√ / √	√ / √
BC (7)	√ / √	0 / √	√ / √	√ / √
SA (7)	√ / 0	0 / 0	√ / √	√ / √
BM (6)	√ / √	0 / √	√ / √	√ / √
SO (6)	√ / √	√ / √	√ / √	√ / √
HV (6)	√ / √	0 / √	√ / √	√ / √
TP (6)	X / √	0 / 0	X / 0	√ / √

Note. √ refers to a correct response, 0 refers to an incorrect response, and X refers to when there was no data for a students' response

Table 1. Students' Initial Responses Regarding Real and Made up Graphs

Number of Correct Identifications[a]	Middle School Students (n = 7)	High School Students (n = 15)	All Students (n = 22)
0	0	0	0
1	0	0	0
2	0	5	5
3	5	3	8
4	2	7	9

Note. Of the 24 students interviewed, only 22 respondents are represented in this table. The other two students' initial responses weren't clear.

[a]In order to receive a score of 4 correct identifications, a student would have had to correctly identify Graph B and Graph D as made up and Graph A and Graph C as real.

Table 2. Students' Final Responses Regarding Real and Made Up Graphs

Number of Correct Identifications	Middle School Students (n = 8)	High School Students (n = 16)	All Students (n = 24)
0	0	0	0
1	0	0	0
2	2	9	11
3	0	0	0
4	6	7	13

Table 3. Faction of Students with Initial Correct Identifications of Graphs

Graph Name	Middle School Students	High School Students	All Students
A	7 / 7	9 / 15	16 / 22
B	2 / 7	9 / 15	11 / 22
C	7 / 7	14 / 15	21 / 22
D	7 / 7	15 / 15	22 / 22

For example, 14 / 15 means that 14 of the 15 high school interviewees correctly identified Graph C as being real in the interview.

Table 4. Fraction of Students with Initial Correct Identifications of Graphs

Graph Name	Middle School Students	High School Students	All Students
A	7 / 8	9 / 16	16 / 24
B	6 / 8	8 / 16	14 / 24
C	7 / 8	14 / 16	21 / 24
D	8 / 8	15 / 16	23 / 24

Table 5. Frequency (percent) of students with responses in each category.

	Specific Data Points	**Variation**	**Center**	**Distribution**	**Informal Inference**	**Context**
Middle School n = 8	4 (50)	5 (62.5)	5 (62.5)	1 (12.5)	3(37.5)	4 (50)
High School n = 16	4 (25)	12 (75)	9 (56.3)	3 (18.8)	7 (43.8)	4 (25)
Overall n = 24	8 (33.3)	17 (70.8)	14(58.3)	4 (16.7)	10 (41.7)	8 (33.3)

Table 6. Frequency (percent) of students responding Agree/Disagree/Ambivalent* that wait-times are the same accompanied by their respective reasoning.

	Agree (A), Disagree (D), AmBivalent (B)	**Specific Data Points**	**Variation**	**Center**	**Distribution**	**Informal Inference**	**Context**
Middle School n = 8	A = 0 (0)	0	0	0	0	0	0
	D = 7 (87.5)	2	2	1	2	2	1
	B = 1 (12.5)	0	1	1	0	1	0
High School n = 16	A = 3 (18.8)	1	0	3	0	2	0
	D = 11 (68.8)	4	10	1	2	4	0
	B = 2 (12.5)	0	1	1	1	0	2
Overall n= 24	A = 3 (12.5)	1	0	3	0	2	0
	D = 18 (75)	6	12	2	4	6	1
	B = 3 (12.5)	0	2	2	1	1	2

*Students unable or unwilling to decide were coded ambivalent (B).

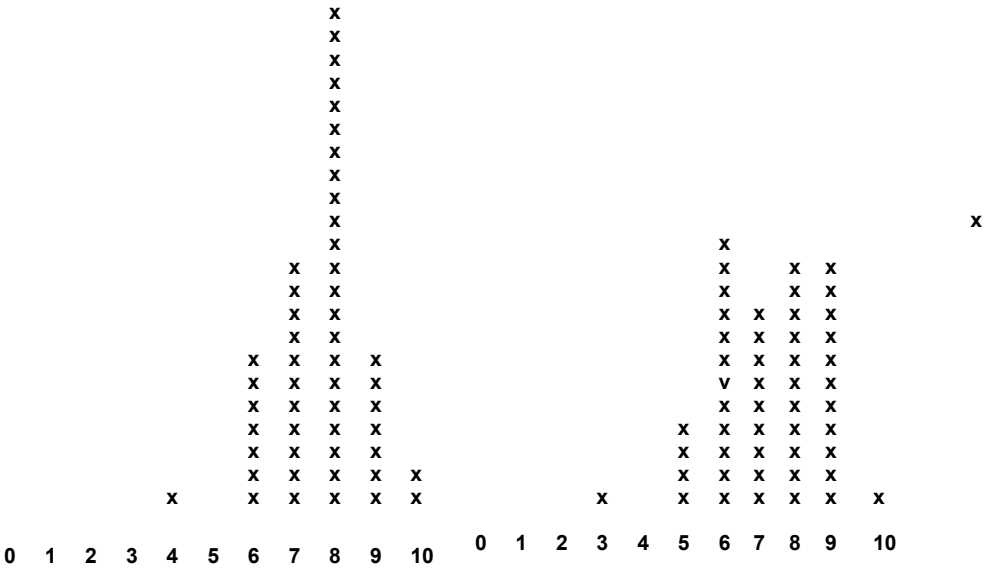
Table 7. Frequency (percent) of students choosing Maximum theaters/Royal theaters/Ambivalent* accompanied by their respective reasoning.

	Maximum (M), Royal (R), amBivalent (B)	**Specific Data Points**	**Variation**	**Center**	**Distribution**	**Informal Inference**	**Context**
Middle School n = 8	M = 2 (25)	0	0	0	0	2	0
	R = 6 (75)	1	2	0	0	5	2
	B = 0 (0)	0	0	0	0	0	0
High School n = 16	M = 3 (18.8)	1	0	0	0	3	0
	R = 11 (68.8)	2	4	0	0	9	2
	B = 2 (12.5)	0	0	0	0	2	1
Overall n = 24	M = 5 (20.8)	1	0	0	0	5	0
	R = 17 (70.8)	3	6	0	0	14	4
	B = 2 (8.3)	0	0	0	0	2	1

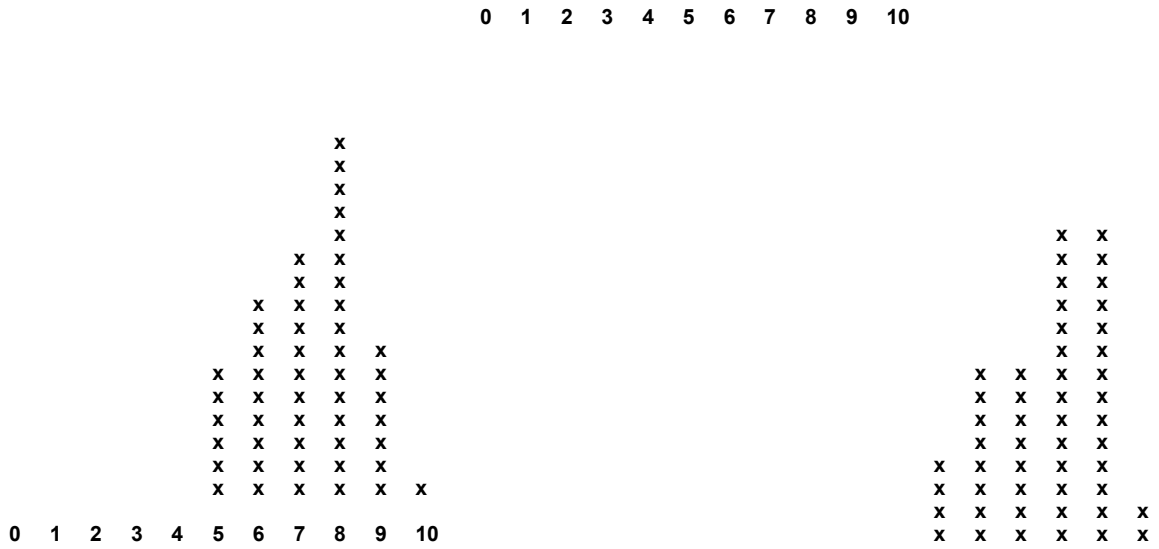
*Students unable or unwilling to decide were coded ambivalent (B).

Working in small groups, students in a class pull samples of 10 candies from a jar that has 1000 candies. They do this 50 times. The jar has 250 yellow and 750 red candies in it. Each time they put the sample back and remix the jar.

Here are some of their actual graphs for the numbers of reds in their 50 handfuls of 10 candies. What do you think about these graphs?



(Graph A) (Graph B)



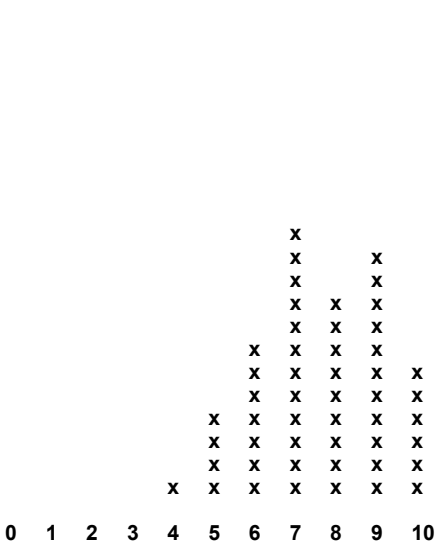
(Graph C) (Graph D)

Figure 1. Four actual distributions

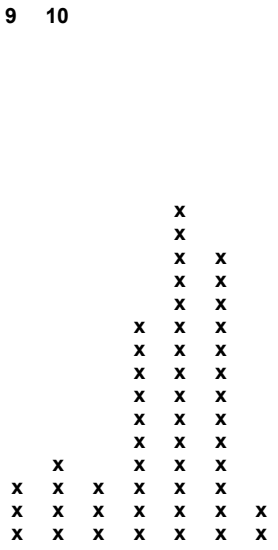
Another class did the same experiment, pulling 50 samples of 10 candies from a jar with 750 reds and 250 yellows, and recording the number of reds. However, in this class some of the groups 'cheated' and didn't really do the experiment, they just made up a graph. Here are some of the students' graphs from that class.

a) Which of these graphs are real? _____and which were made up_____?

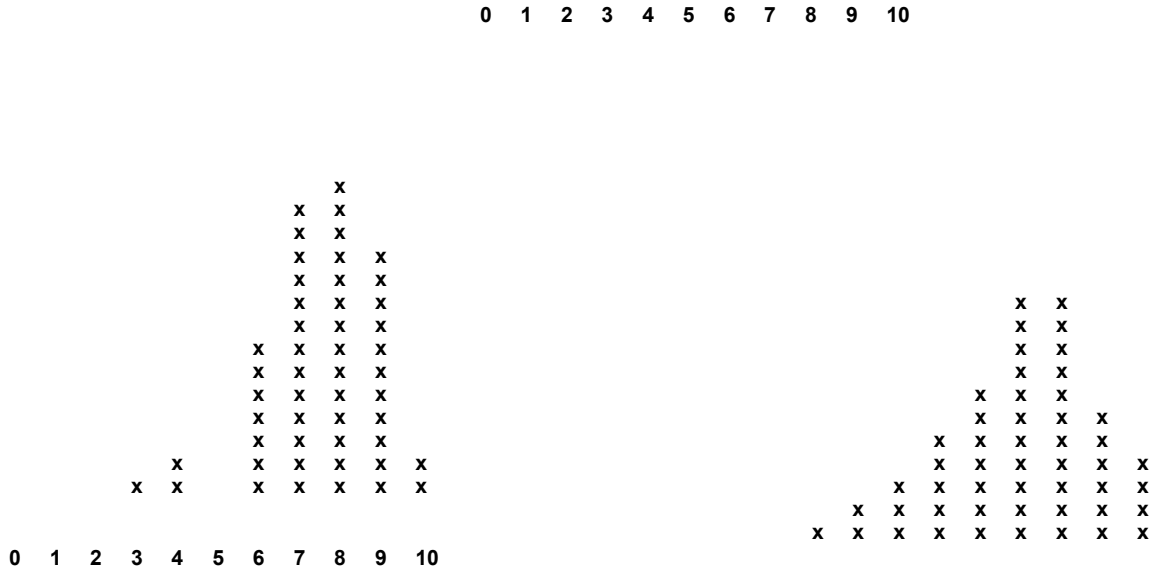
Why do you think this? (Graphs from samples from a 750 red—250 yellow mixture)



(Graph A)



(Graph B)



(Graph C) (Graph D)

Figure 2 Which distributions are real and which are fake?

Movie Waiting Time. A recent trend in movie theaters is to show commercials along with previews before the movie begins. The *wait-time* for a movie is the difference between the advertised start time (like in the paper) and the ACTUAL start time for the movie.
A class of 21 students investigates the wait-times at two popular movie theater chains: Maximum Movie Theaters and Royal Movie Theaters. Each student attended two movies,a different movie in each theater, and recorded the wait-tiomes in minutes below.

Maximum Movie Theaters:

5.0	12.0	13.0	5.5	9.5	13.0	5.5
11.5	8.0	8.5	14.0	13.0	8.5	7.0
8.5	12.5	13.5	11.5	9.0	10.0	11.0

Mean = 10 minutes; Median = 10 minutes

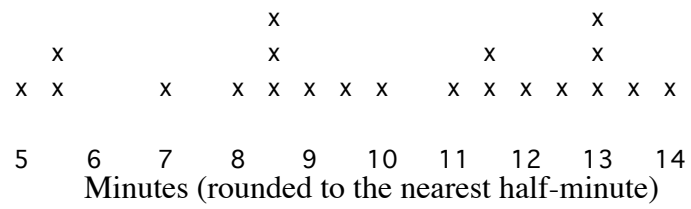
Royal Movie Theaters:

11.5	11.0	9.0	10.5	8.5	11.0	9.0
10.5	9.5	8.5	10.0	11.5	10.5	8.5
9.0	11.0	11.0	9.5	10.0	9.0	11.0

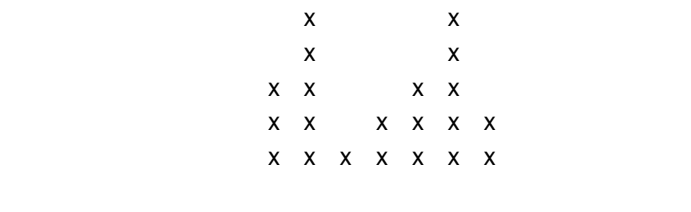
Mean = 10 minutes; Median = 10 minutes

Wait-Time for Movies

Maximum Movie Theaters



Royal Movie Theaters



a) a)What can you conclude about the wait-times for the two theaters?
b) One student in the class argues that there is really no difference in wait-times for movies in both theaters, since the averages are the same. Do you agree or disagree? Why?
c) Which of these theater chains would you choose to see a movie in? Why?

Figure 3. Movie Wait Time Task