Reflecting on Students' Understanding of Data

Kay McClain

Ongoing discussions about students' experiences with data analysis include debates about the role of graphical representations in supporting students' understandings (Shaughnessy 1992; Lehrer and Romberg, in press; Hancock, Kaput and Goldsmith 1992). More specifically, do students first need to know how to construct various types of graphs before they can engage in an analysis of data, or can they learn to construct various types of graphs by engaging in data analysis? Further, can they engage in data analysis before they have acquired a conceptual understanding of the multiple forms of data representation? These discussions also include a debate over the role of representative values, such as the mean, mode and median (Mokros and Russell 1995). What are the students' understandings of representative values? How should representative values be introduced? What role does students' understanding of representative values play in their ability to analyze sets of data?

This article is intended to address these issues by discussing what happened as I observed several groups of Grade 7 students working on performance tasks designed to assess their understanding of both graphical representations and representative values. During my interactions with these students, their activity made me reflect on my then-current beliefs about what it means to know and do "data analysis" in the middle grades. In particular, as I observed students attempting to find a way to "represent" a data set by debating the advantages and disadvantages of the graph as a visual, single-glance impression of the entire data set versus the mean as a single-number numerical summary, I began to rethink what might be involved in an instructional sequence that addresses these concepts. By taking an in-depth look at the way that students reasoned about representing a set of data and conveying my reflections on their processes, this article highlights the importance of teachers' supporting students' development of conceptual understandings of multiple forms of data representation and representative values in the context of ongoing data analysis.

Classroom Episode

A shift in the way I thought about students' experiences with data occurred as I was working with Grade 7 students on several performance-assessment tasks. The tasks were intended to yield information about how the students reasoned about a set of data and were designed to be accomplished in group settings. Specifically, I wanted to try to understand how the students might go about organizing and subsequently representing a set of data to generate a summary of the information. This preassessment information would, in turn, be used to guide the research team's decisions about instructional tasks that would be introduced in a unit on exploratory analysis. As a result of the need to clearly understand students' interpretation of and reasoning about the task, two of the group discussions and the subsequent whole class discussion were videotaped. Viewing the tape gave me the opportunity to analyze the students' ways of reasoning beyond what I observed as I monitored the different groups.

One task asked students to summarize the results of a hypothetical survey to create a report for the principal and parents. The survey results included the number of hours of television that 30 Grade 7 students watched in one week. The task is shown in Figure 1. In anticipating how the students might begin to reason about organizing the data, it seemed obvious to me that a histogram would be a clear and concise way to represent the information. For me, the histogram would preserve the variation in the data set while simultaneously giving a holistic impression of the trends and patterns within the data set without the need to digest each piece. The use of the mean for this particular set of data would appear to oversimplify the information, eliminating the nature of the variation. However, as this task was not designed for preassessment purposes, I was not sure how ready-to-hand the use of histograms was for the students. Further, my assessment of their performance would not be based solely on whether they made a histogram and made it correctly but would focus more on how they reasoned about organizing and representing the data. My colleagues and I needed this type of information to guide the design of the instructional tasks for the unit on exploratory data analysis.

As students began working in their groups, I walked around the classroom to try to monitor their activity and begin to understand how they were reasoning. Several groups began by finding the mean of the data set. One group asked for the calculators, and it could be argued that their request triggered a request from the other groups. Nonetheless, several groups began to methodically calculate the mean of the 30 responses.

As I continued monitoring the groups, I noticed that of the groups that found the mean, some of them subsequently rejected it as inappropriate. They reasoned that just reporting a mean of 10.56 for this set of data was insufficient, arguing that just one number did not provide enough information for this particular data set because "a bunch of the numbers were way above and a bunch of the numbers were way below" the mean. They then proceeded to discuss how to make a graph that would better represent the data, keeping all the features visible. However, for one group, the use of the mean became an intense topic of discussion. In particular, Amee and Latisha argued that you cannot use the mean, whereas Tony insisted that the assignment was to tell about the 30 students "all together," which was exactly what the mean did. Tony seemed to recognize the need to take the variation into account, and he thought that the mean, as an arithmetic average, did just that.

Amee: You cannot average it out, Tony.
Latisha: You cannot average it.
Tony: If I want to, I can.
Latisha: Listen, Tony . . . .
Tony: [Interrupts] They said all together, *all together* now.

Latisha and Amee then tried to explain to Tony that some of the students in the survey only watched 1.5 hours, so the average is "way off."

Amee: Tony, when you average it out, it is supposed to come somewhere close to [1.5].
Tony: It's not supposed to come out close to 1.5.
Amee: But you cannot do [the average].
Tony: I can. You really can, it just might not be accurate.
Amee: It's way off, it's *very way off*. It's so off, you cannot use the answer.
Tony: Yes, I can.
Amee: Well, you use the answer, but I'm not going to use that answer.

I found it interesting to note that this discussion appeared to focus on whether you could actually use the mean, not whether it would provide a clear representation in this particular task. I interpreted the students' discussion as suggesting that in certain instances you can use the mean and in others it simply does not work. However, their inability to clarify their true understandings was problematic for me. As a result, I intervened in the conversation to help clarify my understanding of their reasoning.

Teacher: Let her give her argument, and then you can give yours.
Latisha: Tony, when I averaged it, it came out to 10. This person here, this is 23 hours.
Tony: So? They said all together, *all together*, not just one person by hisself [*sic*].
Teacher: [Latisha's] agreeing with you that [10.56] is the average. What she is saying is that she doesn't think it is a good way to tell the story of these numbers because there is so much variation here. Because if you say 10, then people might think that everybody watches 10. Some don't even watch any.
Nathan: See, you don't watch 10, you watch about 20.
Tony: So, y'all are just picturing one guy by himself.
Amee: That's what you are doing.
Tony: No, I'm not.

Figure 1. Survey Data Given to Students

How Much Television?

Below are the results of a survey taken of 30 Grade 7 students to find out how many hours of television they watch in a week. The principal has asked you to summarize and represent these data in some form so that parents will be able to understand them quickly when they are posted on the bulletin board. The principal also asks you to write a short report for parents, explaining what the data show.

1.5	21	12.5
0	2.5	15
23	19	4
14	8	16
13.5	16.5	6
4.5	9	18
5	10.5	8.5
6	3	9
11.5	3.5	19.5
13	10	9

At this point, I was still unclear whether Nathan, Latisha and Amee thought that the variability in the data set resulted in the mean's inability to give a complete report of the situation in this particular instance or whether they thought that since some of the data points differed so markedly from the mean that therefore it did not work. I clearly needed to understand those two very different interpretations of the mean to plan how to proceed with further instructional tasks.

As a result of the groups's ability to reconcile their differing interpretations, Tony decided to use the mean and others in the group decided to make a graph. They began by discussing the possibility of grouping the data. Initially they discussed grouping the data from 0 to 10, from 10 to 20, and from 20 to 30. However, Nathan pointed out that the data did not go to 30, so they charged the upper bound to 25. I found this decision intriguing, as the largest data point was not 25 but 23, and changing the upper bound from 30 to 25 would in no way affect the height of the bar. It would, however, create unequal data intervals. The fact that their intervals were not of the same size was not problematic for the students, nor was it represented in the width of the bars in their graph (see Figure 2 for the final version). However, since I was anticipating students' decision that this set of data could best be represented by a histogram, the issue of inconsistent intervals was problematic for me. Further, I was not clear whether the students were making a modified histogram or simply grouping the data points into categories that they named with numeric intervals.

I chose not to intervene in their discussion other than to ask clarifying questions. At this point, I hoped that the issue of inconsistent intervals would arise during whole-class discussions. I judged that it would be much more productive for all the students to be engaged in a discussion about this issue than for me simply to "correct" their mistake. My telling the students would not constitute a basis for their understanding.

As the groups finished their investigations, I asked each group to come to the chalkboard and present the results of their work. The first group to report was Latisha, Nathan and Amee. When they finished, other students in the class appeared to accept their graph as a reasonable way to proceed and a very significant way to represent the data. Interestingly, in creating the graph, they modified their intervals so that they now ranged from 0 to 10, 11 to 20, and 21 to 25. This change occurred as they were placing data into three categories. When I asked where they placed 10.5, they explained that they had rounded each data point to the nearest whole number.

At this juncture, I did not point out the problem with rounding the data and the effect that it has on the representation. Further, I did not see it as my role to ensure that the students made the graphs "correctly" or used the representation that I had envisioned. Instead, I was much more interested in ascertaining what the students did know and thinking about how to use their current understandings as building blocks for later lessons. If these issues did not emerge as problematic for the students, I would be forced to impose conventions, such as how to "correctly" make the graph. Had I simply corrected what I perceived as a mistake, subsequent activities might have caused the students to focus on guessing what I wanted instead of allowing them the opportunity to reason independently. Further, I was still unclear whether the students viewed the data along a continuum of values or had simply categorized the data points. My imposing a continuous scale, as was necessary for the histogram, would be very problematic if the students, at this point, were not reasoning in that manner. For me, a better alternative was to use my knowledge of their current understandings to try to structure subsequent tasks that would make these issues the focus of the investigation.

The second group to share its results drew what I would classify as a modified histogram (see Figure 3).

Figure 2. Group 1's Graph

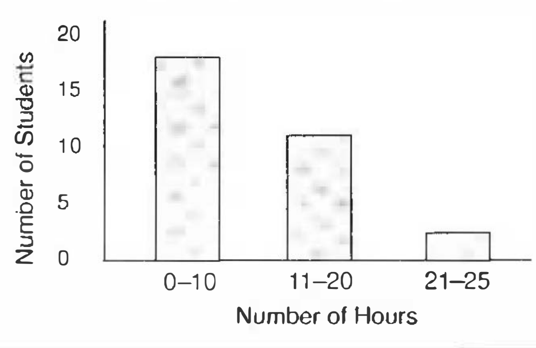
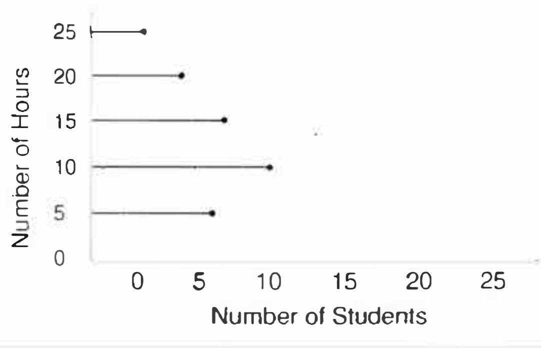


Figure 3. Group 2's graph



The members grouped their data into intervals with a range of five, placing the bar at the upper limit of the range of the interval. In addition, instead of making clearly defined bars, they used line segments. Further, in constructing their graph, they had used a large dot to mark the endpoint of the segment. Therefore, the length of the segment represented the number of data points that fell in the interval (that is, from 0 to 5, from 6 to 10 and so on).

The dot became an issue for other students in the class when I asked, "How are these two graphs alike, and how are they different?"

Andy: That's [points to group 2's graph] a line graph and that's [points to group 1's graph] a bar graph.
Teacher: That's a line graph, and this is a bar graph. Anything else?
Carla: I thought it was supposed to be bars and not like little lines . . . like bars.
Teacher: Why do you think it is bars?
Carla: Because if you do a line, it is supposed to go up at the time when it goes up, and it goes down when . . . it goes up when the rate is high and goes low. . . .
Lynn: I think it is supposed to have bars when it goes vertical [*sic*] like that.
Teacher: What if she is calling these skinny bars? What if she is saying that these are really just skinny bars?
Maggie: With dots on the ends of them?
Jose: With dots on them? I mean, you could do that, but you wouldn't have a line on them, you would just have the dots.
Paul: A line graph is supposed to be connected to other line segments.

For the students, certain conventions were associated with making graphs, and if you used dots, then you must be making a line graph of connected dots. The rules for the use of dots in making graphs seemed very clear to the students; however, they appeared to take great liberties with what I had interpreted as histograms. Their notations of "school mathematics" became intertwined with their goal of making a representation of the data.

The third group to present their graph also drew what I would call a version of a histogram. They appeared very clear on the rule that the intervals all had to be of the same size. However, in their "histogram" the intervals were each composed of 6 of the 30 data points. They said that they decided to use six data points in each interval because you could divide 6 evenly into 30 to get five groups. They had ranked the data from least number of hours of television watched to the greatest, as if finding the median, and then divided the ranked data into five groups of six numbers each (see Figure 4a). They then totaled the six data points in each group, which gave them the height of the respective bar (see Figure 4b). The result obviously gave a series of bars with increasing heights, since the subsequent bars contained the larger data points (see Figure 4c). The group members labeled the horizontal axis according to which 6 of the 30 data points were contained in the interval, and they used the vertical axis as a scale for the total hours for each of the six groups of data points.

After group 3 finished explaining its graph, I again asked how the three graphs were alike and different. The students' discussions tended to focus on the direction of the bars (that is, vertical or horizontal) and the labeling of the axes. For instance, many students noted that group 2 and group 3 both labeled the horizontal axis as "number of students." The fact that the number represented very different reports was of no consequence. They also noted that group 2 was the only group whose graph was "sideways." Their focus was on the superficial features of the graphs instead of the underlying meaning and intent.

The members of the next group to report stated that they had calculated the mean. I then asked why they thought that the mean was a good way to represent the data.

Mark: Well, we averaged it out, and it worked pretty good; well, because that's like saying, well, when you make a 100 on something and 60 on something and they average that out, like on your report card, that's really not right because you made a high grade on the one thing and a low grade on another.

It appeared that in Mark's justification for the use of the mean, he was questioning whether it was "right," since he related it to finding an average grade, he would not want to use the mean if he had a high score and a low score. I found it intriguing that, as he offered his justification to the class, he appeared to be reconceptualizing his understanding of the mean as a way to represent his group's data. Of the remaining three groups, two calculated the mean and the third made a bar graph of each value in the data set.

As the class drew to a close, I realized that I had generated more questions for myself than answers. The students had obviously acquired ways to reason about the data. However, their ability to make judgments tended to be impaired by their notions about "rules," for example, if you use dots, they have to be connected. Also, group 3 was certain that each bar had to represent the same number of data points, but the intent of the graph was never clarified. Initially, I had been concemed not about whether the students could make the graphs correctly but with how they reasoned about the data set. However, as a result of interacting with the students on multiple tasks, I realized that their previous experiences with learning how to make graphs would clearly have a large impact on future instructional goals. Apparently, the set of rules associated with the construction of graphs formed the basis of their initial investigations with graphs. They perceived that I was now asking them to remember all the rules as they made reasonable judgments about the data sets rather than to reason and construct their own ideas. Asking them to create their own representation appeared to be a shift in the type of activity that they normally associated with sets of data. Also, their interpretation of the mean as a representative value was unclear to me. Students clearly knew how to calculate the mean, but their ability to talk about it as a reasonable way to represent data set seemed limited.

Implications for Instructional Tasks

The results of my observations of the students' ways of reasoning on this and other tasks had serious implications for instructional "next steps." To support students' development of ways to reason logically about data, instructional tasks would need to build from the students' current understandings. Simply restating rules and procedures for the proper ways to construct types of graphs would not equip the students with powerful tools for data exploration. Their subsequent activity would be reduced to deciding which graph to use for which set of data according to some predetermined criteria. Further, problematic areas in their current understandings would need to be highlighted in the context of investigations so that those areas could then become the focus of discussion. This outcome would require finding situations in which the data presented the opportunity for students to reason about the implications of their decisions.

In working with my colleagues to develop a sequence of specific tasks that would capitalize on the students' current understandings, we decided to focus on the use of the mean as a representative value. We thought that this understanding could be best achieved by having students work to compare two sets of data to make a decision. In this situation, each choice would result in a consequence. We hoped that students would then come to see the value and limitations of the mean. As we deliberated, we decided to find situations in which the two data sets had very similar means even though the individual data points in one of the sets varied greatly. In these particular situations, it might be important to know more than just the mean; one might also need to know the consistency, or the range, of each set. Students would then need to work to find another way to tell the whole story of the data.

One such investigation involved comparing the hours of use of a sample of 10 separate batteries from each of two different brands of batteries. Students were immediately engaged, since most of them depended on batteries for some type of personal electronic apparatus. As they began to analyze the two sets of data, their investigation led to discussions about what brand had "more bad batteries" and which brand was "more dependable." During the discussion, some students argued that if you used the battery with the higher average, over time that brand would be the better

Figure 4. Group 3's Data Points, Rankings and Graph

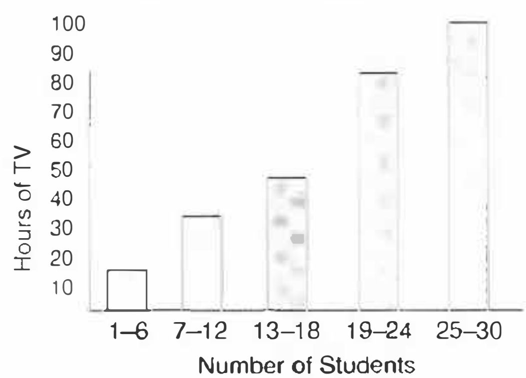
(a)
Ranked data separated into five groups

0 1.5 2.5 3 3.5 4 / 4.5 5 6 6 8 8.5
/ 9 9 9 10 10.5 11.5 / 12.5 13 13.5 14 15 16
/ 16.5 18 19 19.5 21 23

(b)
Totals for the five groups of data

0	1.5	2.5	3	3.5	4	= 14.5
4.5	5	6	6	8	8.5	= 38
9	9	9	10	10.5	11.5	= 59
12.5	13	13.5	14	15	16	= 84
16.5	18	19	19.5	21	23	= 117

(c)
Group 3's "histogram"



choice. The students then worked to find other ways to represent the information to support their arguments. This type of discussion provided opportunities to build from students' investigations to support conceptual understanding of the mean in relation to a set of data and of the mean in relation to a graph. The development of this type of task clearly requires that we find situations that not only appeal to the students but also feature data that fit with our pedagogical agenda. This approach is far more work-intensive than teaching from a textbook. However, I believe it to be both necessary and appropriate.

Conclusion

As a result of my interactions with these students, I have further refined my ideas about how to support students' development of statistical reasoning. I believe that students need to develop ways to reason logically about data—not memorize rules and procedures. For me, this belief involves students in tasks that permit the exploration of multiple forms of representation. Students' experiences with data, graphs and representative values need to reflect the expectation that they reason about situations rather than simply apply rules and algorithms. For instance, students should use their analysis as the basis of arguments. By making reasoning the focus of our instruction, we then change our expectations. We focus on the meanings that graphs and representative values have for the students in light of their investigations.

In the past, the distinction I often made between "graphing" and "finding the average" caused students to infer that these two types of representations are not related. They therefore lacked the understanding that both representations are models of the data set, each highlighting different aspects of the set. By allowing students' initial exposure to data analysis to be in the context of exploration of data sets, we give purpose to the investigations and subsequent representations. The specifics of constructing graphs can emerge from the importance of clarifying the representation. For instance, the graph presented by group 1 offered an opportunity to build on students' contributions while clarifying the purpose of a histogram. In this setting, the problem of rounding data points could be highlighted as students wrestled with the notion of a continuous scale.

As a result of thinking differently about how graphical representations should be incorporated in the middle school curriculum, I see the value in allowing students to engage in analyzing data before they have mastered all the conventional "tools." In this way, the tools emerge from students' investigations, thereby acquiring meaning in the context of their activity.

References

Hancock, C., J. Kaput and L. Goldsmith. "Authentic Inquiry with Data: Critical Barrier to Classroom Implementation." *Educational Psychologist* 27 (1992): 337–64.

Lehrer, R., and T. Romberg. *Children's Constructions of Data, Chance, and Inference in Design Contexts*, in press.

Mokros, J., and S. J. Russell. "Children's Concepts of Average and Representativeness." *Journal for Research in Mathematics Education* 26 (January 1995): 20–39.

Shaughnessy, M. "Research in Probability and Statistics: Reflections and Directions." In *Handbook of Research on Mathematics Teaching and Learning*, edited by D. Grouws, 465–94. New York: Macmillan, 1992.

The research and analysis reported in this article were supported by the Office of Educational Research and Improvement (OERI) under grant number R305A60007 through the National Center for Improving Student Learning and Achievement in Mathematics and Science. The opinions expressed do not necessarily reflect the views of OERI.

The author would like to thank Judi Zawojewski for her comments on an early draft of this article as well as Paul Cobb, Koeno Gravemeijer and other members of the research team for their efforts at analyzing students' work and in designing instructional tasks.

Reprinted with permission from Mathematics Teaching in the Middle School. *Volume 4, Number 6, (March 1999) pages 377–80, an NCTM publication. Minor changes have been made to spelling and punctuation to fit ATA style.*