STUDENTS' USE OF MODAL CLUMPS TO SUMMARIZE DATA

Clifford Konold, Amy Robinson, Khalimahtul Khalil,
Alexander Pollatsek, Arnold Well, Rachel Wing, and Susanne Mayr
University of Massachusetts, Amherst, USA

We interviewed 7th and 9th grade students to explore how they summarized and reasoned about data. The students were near the end of an eight-week collaborative research project in which they analyzed data they had collected on the types and frequencies of animals killed on town roads. During our interviews, students worked with data similar to those they had collected to answer questions we posed about conditions that might affect the number of animals struck by cars. To summarize their data, students tended to use a "modal clump," a range of data in the heart of a distribution of values. These clumps appear to allow students to express simultaneously what is average and how variable the data are. Modal clumps may provide useful beginning points for explorations of more formal statistical ideas of center.

INTRODUCTION

Everyday situations often require us to summarize phenomena whose outcomes vary or are inherently uncertain. "It will take about 10 minutes to walk to the diner from work. Lunch will run about $8.00." In such situations, we almost never use formally computed measures. One reason for this is that we rarely need to be precise. But even if we wanted to be, we would ordinarily not have ready access to either the data or the computational tools to compute, for example, a mean value of personal travel times or lunch prices.

We encounter formal averages in disciplines and practices where accuracy is critical and in which claims are subject to independent verification. We also encounter them in classrooms where students are introduced to disciplined inquiry. Such classrooms are fertile sites for investigating opportunities and conflicts that arise as students learn practices that are similar in many ways, but quite different in others, to everyday practices and purposes.

In this study, we explore how young students in school settings reason about variable data and, in particular, how they communicate ideas about typicality and variability of data. This research builds on several recent studies that have reported people summarizing distributions by specifying a range of values in the heart of the distribution (Bakker, 2001; Cobb, 1999; Konold & Higgins, in press). Cobb (1999), for example, described a seventh grade classroom in which students comparing two dot plot distributions of car speeds began comparing the "hills" in the two distributions. Konold and Higgins (in press) described third grade students who used what the students termed a "middle clump" to summarize a distribution of family sizes. Noss, Pozzi, and Hoyles (1999) reported similar observations from their study of a group of practicing nurses reasoning about data they commonly encounter on the job. Our primary motivation for learning more about the ways students choose to summarize data is to improve instruction. By building instruction on what students already know and do, we believe we can help students develop more robust conceptual foundations.

THE INTERVIEWS

We interviewed teams of students at two New Hampshire public schools, Pinkerton Academy, in Derry, and Rye Middle School, in Rye. During the previous 8 weeks, classrooms at each school had been participating in a collaborative science project (Bartlett, 2001). The project, known by the morbid moniker "Roadkill," is one of a new genre of "network science" projects (cf. Feldman, Konold & Coulter, 2000). As part of the Roadkill project, students observe the number and type of animals killed on local roads and share this information over the internet with partner schools. (For a critical review of this and four other network science projects, see Biehler & Schweynoch, 2001.) In both classrooms, students formed small collaborative research teams in which they formulated questions and hypotheses about the type and number of animals that were killed. They then investigated their questions by analyzing appropriate subsets of the project data pool. We chose to interview students in these particular classrooms because of their deep involvement in posing questions, collecting and analyzing data, and reading and discussing information about local wildlife and habitats. We hoped to get a richer picture of how students

reason about data than if we observed students working with data that were unfamiliar or of no particular interest to them.

At Pinkerton Academy we interviewed three teams of two students from a ninth grade science class; at Rye Middle School we interviewed four teams of four students from a seventh grade general science class. The teams had been working together for several weeks prior to our interviews. They thus had established patterns and norms for interacting, delegating, and sharing tasks. Because we did not want to disrupt these patterns during the interviews, we posed questions to the team and let the team decide how to respond to them. We did not require that the team reach a consensus on questions we posed, and we generally refrained from encouraging silent members to speak up. However, we occasionally directed follow up questions and requests for clarification to specific team members, and probed to see what individuals thought of a proposal advanced by another team member.

During the interviews, students worked with data similar to those they had collected during the class project. We posed a series of tasks and questions about conditions that might affect the number of animals struck by cars. We videotaped the interviews, then produced written transcripts and used both of these for our analyses. Our primarily unit of analysis here is the team. We describe responses the team settled on and note disagreements among team members when they occurred.

Q1: HOW MANY DEAD ANIMALS DO YOU TEND TO SEE EACH DAY?

During each interview, students had several opportunities to summarize data. The interview began with the question, "How many dead animals do you tend to see each day" on town roads. Our purpose was to see how students would characterize the data they had been observing over the past few weeks. Would they make use of statistical notions such as center or spread? To avoid directly prompting the use of averages, we were careful in this initial probe not to ask what was "average," "typical," or "usual." One team from Derry (D1) simply pointed out that the number varied daily:

Nick*: Um, some days there weren't, there wasn't too much and then the next day there'd be like 3 or 4 of one thing [kind of animal]. It was kinda of, it wasn't real consistent.*

The other six teams also mentioned that the number varied but, in addition, spontaneously gave an indication of what "usually" happened. For example, team D3 responded:

Maria*: Per day? There were – I don't know. Anywhere from, like, none to 12.*
Cara*: Yeah.*
Maria*: It was usually around like…maybe 3 or 4.*
Cara*: Yeah 3 or 4.*
I*: What do you mean by "usually around 3 or 4?"*
Maria*: Like the – most of the time it was – like sometimes – someday it would be, like there'd be 1 or maybe 2. But some days it would, like as much as 12.*
Cara*: Yeah, for the majority of the days…3 or 4.*

Four of the teams used the term "average" in summarizing their observations, as did this team from Rye (R2):

Grant*: Well, it would change every day, depending on if it was raining, or sunny…*
Ryan*: Or like weather, climate, stuff like that.*
I*: Right.*
Anita*: But on average, probably be like 1 or 2. Not many.*

Students' responses to this initial probe on the whole were consistent with the data their classes had collected. (The daily means were .9 in Rye and 3.8 in Derry.) They tended to use ranges to describe what usually happened, qualifying these with terms such as "around" and "probably." These qualifications emphasized the variability in their data. Shaughnessy, Watson, Moritz, and Reading, (1999) have observed that statistical practice and instruction has tended to focus on measures of center to the exclusion of measures of spread. But clearly these students seemed intent on communicating both these aspects of their data and, if anything, seemed hesitant to use a formal measure of center.

Q2: WHAT WOULD DATA FOR 15 DAYS LOOK LIKE?

After students described what they had observed, we asked them to make up data they might reasonably expect to see over about 15 days, graphing them as a stacked dot plot. Our interest was to see if they would produce data consistent with their summaries. None of the students were familiar with the stacked dot plot. The interviewer (Konold) introduced it by plotting one or two data points on a numbered axis as he verbalized, e.g., "so here's a day on which 2 animals were killed." Then he asked the students collectively to add more values. Each student had a pencil for adding data to the plot. During the remainder of the interview, most students were able to correctly interpret these displays. On the few occasions when students became confused, they mistakenly interpreted a stack of, for example, three x's over the number 7 as representing 3 animals killed on day 7 rather than as 3 days on which 7 animals were killed.

Table 1 shows the data each team generated, which are reasonably consistent with their initial summaries. From what we could determine, their values are also fairly consistent with the data they had actually collected. The two exceptions were teams D1 and R1. Students in team D1 produced a bimodal plot based on their memories that they less frequently observed 2-3 animals on a day than they observed 0-1 and 4-5. We did not have access to their daily values and so could not check this observation independently, but it seems unlikely.

Table 1

Hypothetical Data generated in Response to Question 2

	Team						
	D1	D2	D3	D4	R1	R2	R3
Initial summary		around 5	usually 3-4	avg 5 or 6	avg 2-3 0-3 1or 2	avg 1 or 2 avg 2	avg 2 or 3
# killed/day							
0	xxxx |		xx	x	xxxxx	xxx	x
1	xxx |	xx	xx	xx	xxxxxx	xx	xxx
2	(gray) |		xx	xxx	xxxx	xxxx	xxx
3	(gray) |		xxxx	xxxx		xx	xx
4	xxx |	xx	xxxx	xxxxx		xx	xx
5	xxxx |	xxxx	xxx	xxxxx		x	
6	|	x	xx	xxxxxx			
7	|			xxxxx			
8	|	x	x	xxx		x	
9	|		x	xxx			
10	|	x		xx			
11				xx		x	
12	|		x	x			
13				x			

Note: The data show the number of animals the students might observe on town roads over about 15 days, where each x represents one day. The hypothetical distributions the students made arranged x's along a horizontal rather than vertical axis. At the top of the columns are team responses to Question 1 asking them to summarize the actual data they had collected (some teams offered more than one summary, as noted). Highlighted in gray are the ranges the teams gave when we asked them to summarize their made-up data (Question 3). Team D1 drew lines above their distribution to show how the distribution would look if they added more data.

Team R1 produced data with a range much narrower than the range of the data their class had actually collected. There is some indication, however, that these students misunderstood our instructions of what to plot. In rephrasing the interviewer's request, April responded, "How many animals do you think would be killed in a day? Like, a normal day?" The interviewer missed this subtle but significant reinterpretation of his question and so did not further clarify the task for the students. Consequently, the students appeared to include in their plot only the most commonly occurring values, which they had just reported were in the range of 0-3.

Q3: HOW WOULD YOU SUMMARIZE THIS PLOT?

We next asked students to summarize the made-up data they had plotted for someone who could not see the graph. This task was slightly different from their first task of summarizing, based on memory, the data their class had collected. In this instance, they were to summarize values they all could see, and they could use those values, if they wanted, to compute formal averages. Would these differences change the way they summarized the data? This task gave us some additional information as well, as we could investigate features of the source data that may influence their summaries. All of the teams summarized their made-up data by again specifying a range of values within the distribution. These ranges appear in Table 1 highlighted in gray. Two of the teams, D1 and R2, referred to their summary as an "average." As we suggest below, team D1 may have had the mean or median in mind. But in most cases, the summaries students gave were not formal point averages.

MODAL CLUMPS

In most of the cases, it seemed that students intended the entire central range they gave us as an indicator of what was usual or typical. We refer to such ranges as "modal clumps." In the case of two of the teams, however, it was not clear whether they thought of the entire central range as their summary ("typical values were 2 and 3") or were suggesting that a point average, such as a mean or median, was somewhere within the specified range ("the average is somewhere between 2 and 3). In summarizing their data, team D1 responded:

Doug: *You'd probably average it. I'd say between 3 to 4 animals a day.*
I: *You'd agree with that to be a decent summary? (looking at Nick)*
Nick: *Yeah, around around - there'd be more - the average number would probably be around the 2 or – the 2 through 3, around the 3 area. But more animals being killed in the 5 or 4 and 0 and 1 numbers than in the 2 and 3.*
I: *Uh-huh. So you're saying the average would be 2 or 3?*
Nick: *The average would be in the middle, but that's kind of almost deceptive because there's more animals killed on either side. But they almost even each other out.*

Doug's expression "between 3 to 4" and Nick's "around the 3 area" suggest that these may have been not modal clumps, but rather estimates of the location of a point average. Statisticians looking at these data would probably say something very similar if they had to guess at the location of the mean. Rita in team R3 used similar language, giving the range "2 to 3" and then immediately rephrasing this response to "about 2 or 3." It is certainly plausible that her first statement "2 to 3" was a range within which she thought some point average would be, and that her rephrasing "about 2 or 3" clarified that the average could be either 2 or 3 (but not both).

However, in all the other instances, as in Pat's (D4) statement below, students seemed to view the entire interval, or modal clump, as their summary:

Pat: *Um, this, it's not, it's not like too many. It's not more around 12's, but they're mostly in the range of the middle numbers: 4 through 8.*

Konold and Higgins (in press) suggested that to students, the ideal "average" is a) an actual value in the data set, b) the most frequently occurring value (the mode), c) located midway between the two extremes both in terms of value (the midrange) and order (the median), and d) relatively close to all the other values. The modal clumps that students they studied constructed included many of these ideal attributes. The clumps tended to be located in the middle of the data set and, viewed as a collection, to constitute the most frequently occurring values. The other values in the data set were relatively close to the borders of the clump, certainly closer than those values would be to any given point average.

Table 2 looks more closely at the nature of the modal clumps that students we interviewed used to summarize their hypothetical data. We have omitted from the table the results from team R1 because they apparently plotted only their modal clump. We also omitted D1, because they may have been bracketing a point average, as the dialogue above suggests. The table shows the percentage of data values that fall below, within, and above a teams' modal clump. Each team's modal clump included a higher percentage of data than either of the other two partitions. This is one reason we think the term "modal clump" is an appropriate descriptor.

Furthermore, the modal clumps of 3 of the 5 teams included a majority of the data (more than 50%). For comparison purposes, we also include in Table 2 various standard statistical summaries of the hypothetical data, such as the median and Interquartile range (IQR). We stated earlier that modal clumps may serve the function of summarizing both where the data are centered and how spread out they are. In this regard it is interesting to note how close the modal clumps are to the corresponding interquartile ranges (IQRs) and how close the medians are to the midrange of the modal clumps.

Table 2
Characteristics of Students' Modal Clumps for the Hypothetical Data they Created.

Team	Modal clump	Distribution of values with respect to modal clump			Statistical properties of data			
		% low	% middle	% high	mean	SD	median	IQR
D2	4-6	18	64	18	4.9	2.5	5	4-6
D3	0-6	0	86	14	4.1	2.8	4	2-5
D4	4-8	23	56	21	5.9	3.1	6	4-8
R2	2-3	31	38	31	3.0	2.9	2	1-4
R3	2-3	36	46	18	2.1	1.2	2	1-3

Note: For each team, the table shows the highest and lowest value of the modal clump and the percentage of values in that team's distribution that were below, within, and above this clump. In the right hand columns are statistical properties of the distributions. Note the similarity between each team's modal clump and the corresponding IQR.

In two instances, students expressed a distinction between the summaries they had given and formal averages. Rita in team R3 said:

Rita*: Just like looking at it, you'd say: "Oh, the average is about 7 […]" but if you did the math it'd probably turn out to be like, I don't know, who knows?*

Similarly, Ryan in team R2 first specified that "on an average there's usually like 2 or 3 roadkill, well, around every day." The phrase "usually like 2 or 3" suggested a modal clump. After saying he agreed with Ryan's summary, Grant continued:

Grant*: I'd say that about like 3 got killed. Like, if you had to average it. Three would…*
Ryan*: You'd probably average it out.*
I*: And what do you mean by 'average' when you say "average it out"?*
Ryan*: Add up all the — total roadkill, and then you divide by how many…*
Grant*: You add all the x's together and*
Ryan*: Yeah.*
Grant*: and then you divide by how many.*

Both of these summaries seemed acceptable to the students in R2 and serve for us as a clear contrast between modal clumps ("usually 2 or 3") and point averages ("3 got killed, if you had to average it"). Grant's phrase, "if you had to average it" is consistent with our sense that for many students, formal averages are not their summary of choice.

SUMMARY
For the students we interviewed, "2 to 3 animals a day" seemed to be a more informative summary of what they typically observed than either a mean or median. Indeed when accuracy is not required, modal clumps have some advantages over means or medians: In addition to indicating where the data are centered, they also give some sense of how the data are distributed. As the students we interviewed used them, modal clumps served as descriptors of the location of the majority of the data. There were various indications, however, that these students did not regard their modal clumps as representing the entire group of data. One indication of this is that they did not use these modal clumps later in the interview to make comparisons between two groups. This finding replicates reports of a number of researchers (as summarized in Konold & Pollatsek, in press).
Earlier research investigated difficulties students had in using the mean (Pollatsek, Lima, & Well, 1981; Strauss & Bichler, 1988). Based on their studies of students in grades 4, 6 and 8,

Mokros and Russell (1995) suggested introducing younger students first to the median, delaying instruction on the mean. More recent research has indicated that for younger students the median is no more intuitive an indication of typicality than is the mean (Konold & Higgins, in press.) The research we report here, along with that of Cobb (1999), Bakker (2001), and Konold and Higgins (in press), suggests that the idea of modal clump may provide a more useful beginning point for learning to summarize variable data.

For instruction to build successfully on the idea of modal clumps, students will need to begin to formalize and study them. The need for formalization could be introduced by having students explore communication problems that would result from each student using his or her own ad hoc methods for composing them. This could set the stage for explorations of alternative ways to define modal clumps (e.g., 1/3 - 2/3 of the range; 40th - 60th percentile). We are currently designing data analysis software (Tinkerplots) for the middle school in which students can use and formalize a variety of informal methods, including modal clumps, to summarize and compare distributions. In future research, we will explore how these capabilities might be used during instruction to help students come to view distribution characteristics such as shape, center, and spread as stable features of variable processes (cf. Konold & Pollatsek, in press.)

REFERENCES

Bakker, A. (2001). Symbolizing data into a 'bump'. In M. van den Heuvel-Panhuizen (Ed.). *Proceedings of the Twe25th Conference of the International Group of PME, Vol. 1* (pp. 81-88). Utrecht, The Netherlands: Freudenthal Institute.

Bartlett, B. (2001). The roadkill project. Online at www.edutel.org/roadkill/alt_index.html.

Biehler, R., & Schweynoch, S. (2001). *Data sharing projects: A critical analysis of five projects from the perspective of competencies and opportunities for interactive and exploratory data analysis.* Technical report online, www.mathematik.uni-kassel.de/didaktik/DataSharing/Startdatasharing.html. University of Kassel, Germany.

Cobb, P. (1999). Individual and collective mathematical development: The case of statistical data analysis. *Mathematical Thinking and Learning, 1(1)*, 5-43.

Feldman, A., Konold, C., & Coulter, R. (2000). *Network science, a decade later: The internet and classroom learning.* Mahwah, NJ: Lawrence Erlbaum Associates.

Konold, C., & Higgins, T. (in press). Reasoning about data. In J. Kilpatrick, G. Martin, and D. Schifter (Eds.), *A research companion to NCTM's Standards.* Reston, VA: NCTM.

Konold, C., & Pollatsek, A. (in press). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education.*

Mokros, J., & Russell, S.J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education*, *26(1)*, 20-39.

Noss, R., Pozzi, S., & Hoyles, C. (1999). Touching epistemologies: Meanings of average and variation in nursing practice. *Educational Studies in Mathematics*, *40(1),* 25-51.

Pollatsek, A., Lima, S., & Well, A. (1981). Concept or computation: Students' misconceptions of the mean. *Educational Studies in Mathematics, 12*, 191-204.

Shaughnessy, J. M., Watson, J., Moritz, J., & Reading, C. (1999). School mathematics students' acknowledgment of statistical variation. Paper presented at the 77th annual meeting of the National Council of Teachers of Mathematics, San Francisco.

Strauss, S., & Bichler, E. (1988). The development of children's concepts of the arithmetic average. *Journal for Research in Mathematics Education*, *19 (1),* 64-80.