Chapter 13

Reasoning About Data

Clifford Konold, Scientific Reasoning Research Institute,University of Massachusetts Amherst
Traci L. Higgins, TERC, Cambridge, Massachusetts

Statistics is a rising star in the Grades K–12 mathematics curriculum. Ten years ago precollege students rarely learned about statistics. Now, following recommendations made in *Curriculum and Evaluation Standards for School Mathematics* (National Council of Teachers of Mathematics [NCTM], 1989), statistics is featured prominently in all current mathematics curricula and is popping up in the Grades K–12 science curricula as well.

There are several reasons the field has come to view statistics as important even for kindergartners. Educationally, statistics provides links to other areas of study—science, geography, and history, for example—in which students can apply mathematical ideas to model and reason about real-world situations. The American Statistical Association (1997) has suggested that more needs to be done to develop the "synergy between teaching statistics and teaching basic mathematics concepts" (Section 2, para. 4). At the practical level, knowledge of statistics is a fundamental tool in many careers, and without an understanding of how samples are taken and how data are analyzed and communicated, one cannot effectively participate in most of today's important political debates about the environment, health care, quality of education, and equity. For those who have traditionally been left out of the political process, probably no skill is more important to acquire in the battle for equity than statistical literacy.

Unfortunately, developers of recent curricula have not had years of experience teaching these topics at the pre-college level to draw on in designing their curricula. Furthermore, much of the research describing how students think about and learn these topics is not quite appropriate for their purposes. Because until recently students in North America encountered statistics for the first time at the university, most of the research on statistical reasoning has been conducted with adults and has focused on ideas central to formal statistical inference. Garfield and Ahlgren (1988) and Shaughnessy (1992) provide reviews of this research from the point of view of statistics education. Whereas these ideas are relevant to objectives at the high school level (see Scheaffer, Watkins, & Landwehr, 1998), they do not figure centrally in the curricula for Grades K–8.

In this chapter, we focus primarily on what we have learned more recently from research about how younger students reason about data, concentrating on ideas that begin developing in early elementary school. We therefore do not review the literature related to statistical inference. One reason for not reviewing that literature here is that a reasonable treatment would require us to also review the development of probabilistic thinking (see the next chapter in this volume). But more importantly, core ideas in reasoning about data tend to get shoved to the wings as soon as statistical inference takes the stage. The issues we discuss here, though basic, are still essential to statistical reasoning in the upper grades.

To clarify our point, let us make a distinction between *statistical inference* and *data analysis*. We can think of statistical inference as playing the role of a jury—deciding what we can conclude about a population on the basis of evidence from a sample. Until recently, it had been common practice among social scientists and others who use statistics to collect data and submit them to some test of significance (*t* test, ANOVA, chi square) without ever examining the data in more detail and without looking at the distributions of values for patterns or trends that might lead to additional insight into the question of interest. Once they had conducted the statistical tests, the analysis was essentially over. Indeed, researchers had been warned of the danger of conducting post hoc analyses on questions they had not formulated before collecting the data (Hand, 1998).

Tukey (1977) regarded this exclusive focus on testing hypotheses as shortsighted. He advocated that statisticians could and should do more with data than simply confirm the theories they had formulated prior to collecting data. He encouraged statisticians to take on as well the role of detective, to search among data for interesting and unexpected results. This he described as "exploratory" data analysis (EDA). To assist the researcher in detecting trends and patterns, Tukey developed a variety of methods for displaying data graphically, including the box plot and stem-and-leaf plot. Though he designed many of his techniques for making quick displays with pencil and paper, fast computing and quality computer graphics have actually been a boon to his general approach, making it possible to quickly manipulate and redisplay data in numerous ways. The computer continues to change the way statisticians work (Cobb, 1997).

Although these new developments in data analysis may be a blessing for statisticians, they can be a curse for the educator who is left to figure out how best to teach a field as it is undergoing rapid change. To complicate matters, currently we must assume that students at every grade have had little prior experience with statistics. Therefore, when current research indicates that students at a certain age have difficulty with a particular concept, we do not know what students this age could accomplish if they had years of prior instruction. Thus we find ourselves stitching together a Grades K–12 instructional sequence in statistics and data analysis that, if effective, will probably need overhauling in a few years.

It is all the more important, given this state of flux, that our efforts to teach these topics be guided by an understanding of the core ideas in data analysis. Otherwise we will likely find our students familiar with an array of graphing and data-collection skills but with no ability to reason intelligently about data.

As mentioned previously, many countries, including the United States, have only recently begun to teach statistics and data analysis prior to college; thus research in these areas has not been a priority. We draw heavily in this chapter on the work of Russell, Schifter and Bastable (2002a, 2002b, 2002c), who have compiled as part of course materials for preservice and in-service mathematics teachers a set of cases written by elementary school teachers about their attempts at teaching data analysis. In our opinion, the reflections of these teachers and their descriptions of students' thinking are the richest source to date on children's reasoning about data and on how children's thinking evolves during instruction.

Overview: Data Analysis as an Iterative Process

We can think of data analysis as a four-stage process: (1) ask a question, (2) collect data, (3) analyze data, and (4) form and communicate conclusions (Friel & Bright, 1998). Listing the stages in this order makes sense because it is more or less the sequence in which a research investigation proceeds. It would be hard, after all, to analyze data before collecting them or to form conclusions before doing any analyses. That these are meaningful stages is underscored by the fact that scientists in many fields compose research reports with major sections that correspond to these ordered stages: hypotheses, methods, analyses, results, and conclusions. Similarly, we have organized this chapter into major sections treating, in turn, issues related to data collection, organization, and interpretation.

Real research, however, seldom proceeds in this orderly fashion. One reason is that conscientious researchers often find themselves backtracking. While writing their report, they think of another analysis to do or decide they need to return to the study site to collect more data. But research does not proceed linearly for a more profound reason, which is that these research phases are interdependent (Wild & Pfannkuck, 1999).

Experienced researchers look forward from the beginning. Although they do not analyze data before collecting them, they imagine doing so and make guesses, or hypotheses, about what they will find. They develop and refine their questions and decide what data to collect by thinking ahead to which statistical methods they can use and to the audience they want to convince. Experienced researchers also look backward. When the time comes to analyze the data, they do so from the perspective of their original question, testing the intuitions they started with against what the data reveal. And their questions often evolve and change as they discover unanticipated results in the data (Moore, 1990).

In these respects, data analysis is like a give-and-take conversation between the hunches researchers have about some phenomenon and what the data have to say about those hunches. What researchers find in the data changes their initial understanding, which changes how they look at the data, which changes their understanding, and so forth.

We need to keep this more complex picture of data analysis in mind as we consider what the research literature tells us about students' statistical thinking. Simplistic views can lead to the use of recipe approaches to data analysis and to the treatment of data as numbers only, stripped of context and practical importance. Conversely, staying grounded in the data and attentive to what they have to say keeps the tools of data analysis—the collecting, graphing, and averaging—in their appropriate, subservient role. As one third-grade teacher observed, "Analyzing data is more than just the sum of using data analysis techniques. It's important not to lose sight of what the data themselves have to tell us" (Russell, Schifter, & Bastable, 2002a, case 19).

Turning Observations Into Data

A data investigation usually begins with a question about the real world, for example, "Are children more active than adults?" One of the first challenges is to transform that general question into a statistical one that we can answer with data, for example, "If we put odometers on the feet of a sample of adults and a sample of fourth graders, which group will travel further in a day?" Among other things, the statistical question allows us to develop measurement instruments and data-collection procedures. By analyzing the data, we answer our statistical question, which ideally, but not always, tells us something about the real question we started with.

In learning how to formulate questions and to collect and analyze data to answer them, students must learn to walk two fine lines. First, they must figure out how to make a statistical question specific enough so they can collect relevant data yet make sure that in the process they do not trivialize their question. Second, they must learn to see the data they have created as separate in many ways from the real-world event they observed yet not fall prey to treating data as numbers only. They must maintain a view of data as "numbers in context" (Moore, 1992) and at the same time abstract the data from that context.

"Creating data" may seem an odd phrasing. However, data are not lying around like melons on the ground to gather up and cart off to the table. Turning observations into data involves an explicit process of abstraction, a process more like impressionist painting than snapshot photography (Hancock, Kaput, & Goldsmith, 1992).

Forming a Statistical Question

We collect data to answer a question or solve a real problem. For example, students in the upper grades at a Grades K–8 school believed that the water from the fountains on the third floor, where their classrooms were located, was better than the water from the fountains on the floors below (Rosebery & Warren, 1992). A combined class of seventh and eight graders decided to determine whether the water on different floors of the school building really was different.

Their first task was to reformulate their question as a statistical one: "In a blind taste test of water samples from each of the three floors, which sample will most students prefer?" On the basis of this question, they could develop a plan for data collection. Rosebery and Warren (1992) do not provide the details of this process, but we can presume that students made a number of decisions before collecting data: Whom would they use as tasters? How many tasters should they test? Should tasters drink directly from fountains or from cups? Should the same students taste all three water samples? How should tasters indicate their preference? Such decisions are part of the process of making a general question a statistical one.

Elementary school students can learn a lot about data as they grapple with issues that arise in formulating statistical questions, especially when they anticipate conducting surveys with the questions they design. By thinking about how they would answer a proposed question, students quickly discover not only the range of different responses but also that multiple interpretations of a question are often possible and that the wording of the question matters. In the words of a second grader, "Everyone has to understand your question. If they don't understand your question, everyone will be answering any old way" (Russell et al., 2002a, Case 6).

In Russell et al. (2002a, case 5), a class of fifth-grade students worked in groups to develop questions to guide a data investigation. One group wanted to find out the number of languages spoken by classmates. They suggested asking the following question in a survey: "Do you speak more than one language?" The teacher pushed them to think more carefully about their question:

Teacher: How do we know when someone speaks another language? For example, is knowing how to say, "Where is the bathroom?" in French speaking French?

Student: No. We mean speaking fluently. (Case 5)

This, of course, raises the issue of what is meant by "speaking fluently." The students discussed ways to further refine their question. They were learning that one needs to formulate questions so that respondents interpret the question in the same way.

Often in trying to make a question more precise, students lose track of what they want to know in the first place. For example, second graders Nadia and Keith were interested in finding out from classmates, "How many states have you visited?" (Russell et al., 2002a, Case 6). However, they soon discovered that students they surveyed interpreted "visited" in many different ways. Nadia offered further criteria (described here by her teacher) for defining a visit:

> A visit could only count if you were going to that state for a specific purpose other than simply driving through to reach another destination. Airports could not count. If you stayed with a friend from out of state it could only count if you really, really wanted to see them and you stayed with them for more than a day. (Case 6)

Nevertheless, in their final survey, the question was phrased as follows: "How many states have you ever set foot in?" This wording had apparently been adopted at Keith's prompting because students could answer it with little ambiguity in meaning. However, Nadia was not satisfied. She thought the phrase "set foot in" missed the point. She wanted to know whether students had traveled to, rather than through, a state. In transforming a general question to a statistical one, the problem is not only in wording it so that people will interpret it consistently but also in making sure the question gives you the information you "really, really" want.

Sampling

Part of formulating a statistical question is deciding what population you want to study. Some of the questions we investigate involve collecting information on the entire population of interest (e.g., everyone in a classroom). However, many questions involve populations that would be impossible or impractical to study exhaustively. In these cases, we collect data from only a part of the population—*a sample*. The quality of our study depends heavily on how we obtain this sample. We will probably get a very poor idea of the outcome of a national election by polling, for example, our friends and neighbors—not only because the sample is small but because it is likely biased, that is, not representative of the voting populace. Our friends and neighbors, after all, tend to hold opinions similar to ours. This problem was dramatically demonstrated in the 1936 presidential election (see Gallup, 1972). On the basis of a sample of more than 2 million people, the *Literary Digest* predicted that the Republican candidate, Alfred Landon, would soundly defeat the Democratic candidate, Franklin Roosevelt. From much smaller samples, both Gallup and Roper correctly predicted a Roosevelt victory. The *Literary Digest* had sent questionnaires to addresses they got from lists of owners of telephones and automobiles. Although this was an easy way to get addresses, it tended to exclude poorer segments of the population who at the time did not own phones or cars but did tend to vote the Democratic ticket.

Thus, it is more important that our sample include a representative cross section of the population than that it be large. And the best way to guarantee a representative sample is to draw it randomly. In a random sample, every member of the population has an equal chance of being included. To draw a random sample of 60 students from a large school, we could put the names of all the students in a container, mix them up very well, and blindly draw out 60.

According to research by Jacobs (1999), some students reject the idea of sampling in general "because they drastically underestimate the difficulties of asking everyone in larger surveys, such as surveys of entire states" (p. 245). Others struggle with how information from a sample could give useful information about the whole population.

Understanding the relationship between sample and population requires grasping that X can represent Y without being Y. The majority of the elementary school students studied by Metz (1999) were unwilling to generalize from a sample of crickets they had studied to the larger population of crickets. Among the arguments students gave for not generalizing from their sample were that (a) you can only know about the cases you observe; (b) to characterize a group, you must test every member of that group; and (c) sampling does not work because of variability in the population.

Jacobs (1999) asked fifth graders to evaluate various sampling methods. She used a variety of contexts, including estimating the number of students likely to buy raffle tickets, identifying favorite lunchroom items, and determining which animal students would prefer as a classroom pet. When faced with conflicting results from samples collected in different ways, nearly a third of the students did not differentiate between results produced by biased versus unbiased sampling methods. Rather, they based conclusions on personal experience or said they would ask a trusted authority, such as a teacher or principal, to estimate the outcome. Another 12% evaluated sampling methods on the basis of whether the results fit with their expectations or were decisive. The latter students favored a survey in which, for example, 100% of those sampled said they would buy raffle tickets over a survey that showed a split opinion, because "50-50's not going to decide it for you" (p. 245).

Students who do accept the idea of sampling often favor methods that are biased. Many elementary school students, for example, prefer self-selected samples because they minimize hurt feelings that might result from being excluded from a sample. One upper elementary school student said,

> The people will choose [to participate] if they want to.... Like if they wanted to do the survey, they will, but if they would not want to, they don't have to—so they're not pressuring anybody. (Jacobs, 1999, p. 244)

Schwartz, Goldman, Vye, and Barron (1998) described various biased sampling techniques that sixth graders used to estimate attendance at a particular booth at an upcoming fun fair. Some students suggested they would survey their friends or those they thought likely to visit the booth. They wanted to sample students likely to attend the booth rather than to estimate the distribution of positive and negative responses for the entire school. Other students favored sampling techniques that directly influenced the outcome of the survey. For example, asked to choose a sample to estimate the gender makeup of their school, a quarter of the sixth graders suggested a "fair split" schema, sampling 25 boys and 25 girls. Watson and Moritz (2000) reported similar findings with third- and sixth-grade students.

Some upper elementary school students do select and evaluate sampling methods on the basis of the potential for bias. Roughly one third of the fifth graders Jacobs (1999) studied used expertlike reasoning to evaluate whether various samples would adequately represent the population. The methods they endorsed included randomization and stratified-randomization techniques. For example, one student favored sampling from a population of school students by randomly selecting five girls and five boys from each grade, because

> that way he has a mixture of boys and girls and who are different ages...because sometimes girls and boys can have different opinions on things and also one age might really like something, but an older age might think that was a terrible idea. (p. 244)

Likewise, 40% of sixth graders in a study by Schwartz et al. (1998) proposed sampling methods that avoided obvious bias. A follow-up study indicated that fifth and sixth graders overwhelmingly preferred a stratified or stratified-random sample to a biased sample. However, the students remained somewhat skeptical about using truly random sampling methods. For example, to select a sample of schoolchildren, roughly 60% of the students indicated a preference for selecting the first 60 children in a line over drawing 60 student names from a hat. Indeed, some students worried that randomization might produce a biased sample: "She might pull out [of the hat] all first-grade names" (p. 255). Students may reject randomization procedures precisely because "the selection of population characteristics [is left] up to chance" (p. 256). Thus, when they can, students purposefully select individuals to represent the crucial population characteristics.

It is interesting that students accept randomization in games of chance and see merit in stratification techniques when sampling opinions. What they often struggle with is how these two ideas connect. Schwartz et al. (1998) observe,

> Even though the children could grasp stratifying in a survey setup and randomness in a chance setup, they did not seem to have a grasp of the rationale for taking a random sample in a survey setup. They had not realized that one takes a random sample precisely because one cannot identify and stratify all the population traits that might covary with different opinions. (p. 257)

Differentiating Between the Observed Event and the Data

As we mentioned previously, the process of turning observations of events into data involves simplification and abstraction. Sampling is but one of several stages in this process. To reason intelligently about the data, however, we must keep in mind that the data refer to those more complex, real-world events. If we forget that fact, we begin treating data as numbers only, making no connections back to the context or real-world question.

Other problems arise if we treat the data as if they were the events we observed, failing to distinguish "between the world and a representation of that world" (Lehrer & Romberg, 1996, p. 70). For example, fifth graders in a study by Hancock and his colleagues (1992) collected data to determine which of three cafeteria meals students liked most. The student researchers conducted surveys in the cafeteria on different days, asking students whether they had "bought," "brought," or skipped lunch on that day ("none"). Their plan was to determine menu preferences by comparing the number of students buying versus bringing lunch on a particular day. Their rationale was that because daily menus were published in advance, many students would decide whether to bring a lunch on a particular day depending on whether they liked what was offered. As they began analyzing their data, the student researchers discovered that they had failed to record what meal was served. On the day they recorded the data, a mark under a column on the survey had a clear meaning, a meaning that was gone once they had forgotten the menu of the day.

In Russell et al. (2002a), a kindergarten teacher gave each of her students a bag of M&M's to count. The class created a frequency bar graph using stick-on notes on which each student had recorded the number of M&M's in his or her package. The teacher asked, "What can you tell from this graph?"

Faruh: We eat M&M's.

Rocky: Joy has the most.

Desmond: We know how many I ate.

Tammy: Andrea's is the most...because hers is a bigger number. (Case 16)

The students associated names with values even though the graph they were interpreting did not show who had counted each bag. The students were basing their interpretations on their memories of counting and eating M&M's rather than on the data they had abstracted from that event.

For many students, data serve merely as pointers to the more complex event (Konold, Higgins, & Russell, 2000). In forming conclusions, they draw without awareness on their memories of the event as well as on the objectified data. As a way to help her students begin to distinguish the information in the coded data from what they knew from observing the events, the teacher in the foregoing episode suggested to her students that they "pretend that the principal walks into our room and looks at this chart, what would he know from this chart?" (Russell et al., 2002a, Case 16.)

In keeping with their view of data as pointers, we see many examples in Russell et al. (2002a) of children producing elaborate drawings that show as much about the events as possible. Their iconic representations, often called *pictographs*, can be difficult and time-consuming to draw and typically include detail that would seem to convey no useful information, what Tufte (1983) referred to as "chartjunk." However, whereas to our eye pictographs may be distracting and unnecessary, for students they may help establish explicit links between the data and the event, possibly helping them reason about the data in the appropriate context. Furthermore, the level of abstraction appropriate for a particular representation depends on the questions students have asked. Given that younger students are drawn to questions about who has the most and where they personally fall within a range of values, it makes sense that their representations would make it easy for them to read off individual values and to identify to whom these values belong.

One indication that students are separating data from events is the emerging ability to reason about data in ways they had not anticipated before collecting the data. In their classroom study of fifth graders, Lehrer and Romberg (1996) worked with students to design a survey of student interests. Among other things, the survey asked students to list their favorite school subject, list their favorite winter sport, and estimate the hours they spent watching TV. After collecting the surveys, the instructors asked the student researchers to come up with questions they could "ask about the data." To these fifth graders, this request was ridiculous. Questions, they countered, could be posed to people, but certainly not to data. The instructors prompted the students with various examples, such as "Which is the least-favorite school subject?" The students successfully used these examples to generate a list of similar questions. But they needed further assistance to see that they could answer many of these questions by analyzing the data they already had. Their initial impulse was to conduct another survey using these new questions.

As Hancock and his colleagues (1992) pointed out, once recorded, data become objects in their own right, objects that we can manipulate and query quite independently of the real-world events they model. We can organize data by stacking, grouping, and ordering cases, operations that would be difficult or impossible to perform on the real-world events themselves.

Because students can reorganize the data in a number of different ways, they can pose and answer questions that may not have occurred to them before collecting the data. For example, kindergarten and first-grade students worked with data gathered from a school lunch count (Russell et al., 2002a, Case 1). While gathering data, students became concerned that although a total of 18 students were in their class, they had recorded only 15 data values. The three missing students turned out to have been absent that day. One student was intrigued that they had gotten an attendance count from data they had gathered about lunches. She wondered how this was possible:

> Can the clothespins [markers used to record answers to the survey questions] tell only one thing? If the clothespins tell us how many school lunches and how many home lunches there are, can they tell us how many are at school—I mean at the same time? (Case 1)

This student understood that the sum of the "yes" and "no" counts would equal the number of students in the class that day. But she struggled with the idea that "fifteen could stand for how many in school as well as how many students were getting lunch" (Russell et al., 2002, Case 1). We see her struggle as the beginning of the discovery that data, once recorded, have a life of their own and that in examining them, new questions may arise that the data can answer.

Organizing and Displaying Data

In this section, we examine issues that arise as students organize data into tables and plots. We continue to emphasize the important role that questions play in determining the representations students create and the information they cull from those representations.

Creating Useful Representations of Data

Data are usually somewhat useless in the form we first collect them. A stack of completed questionnaires is like a messy room in need of a good cleaning. To find what we want, we must organize the information, and how we organize the information, or data, depends on what we want to know (Biehler, 1989; Cleveland, 1993).

There are no fixed criteria for judging one data display as superior to another. Rather, the relative value of a plot depends on its intended purpose. Thus, a plot with labeled axes is not necessarily better than one without labeled axes. On the one hand, suppose students wanted to quickly make a frequency bar graph to help them determine where the data were centered. It would be unnecessary in this case to label the axes. In fact, taking the time to do so or to add other fine detail to the graph would be squandering time that could be spent thinking about their question. On the other hand, if these same students made a graph to communicate their findings to the whole class, then labeling the axes and taking care to make the graph easily readable would probably be crucial to achieving their goal.

Likewise, one *type* of representation is not inherently superior to another. Graphs are not better than tables; bar graphs are not better than pictographs. Each of these representations is good for some purposes and not as good for others. Sconiers (1999) described a project undertaken by a kindergarten class that frequently asked their teacher for help tying shoes. The class reasoned that if they all knew which of them could tie shoes, those who did not know how could get help from those who did. After conducting a survey, the students posted a list of names of classmates who could tie shoes. Had the class not been trying to solve that specific problem, they might have made a graph showing how many students could and could not tie shoes, but that plot would have been useless for their purposes. The list worked.

Although no paticular type of representation is inherently better than another, some representations are harder than others to learn to interpret (Bright & Friel, 1998). Roth and Bowen (1994) argued that as we represent numerical data with maps, lists, graphs, and equations, we move from concrete to increasingly abstract statistical representations. As we move along this progression, information about individual data values becomes obscured, disappearing into larger aggregates. According to Feldman, Konold, and Coulter (2000), increasing the level of aggregation allows one to

> perceive ever more general features of the data at the expense of being able to identify individual data values. One can easily forget, however, the learning required to interpret the more abstract statistical plots. As a result, [educators] often encourage students to use plots and summaries before they sufficiently understand them and, by doing so, effectively pull the rug from beneath them. (p. 119)

As Bright and Friel (1998) pointed out, graphs also use axes and other display elements in a variety of ways. We can see some of these differences in the two plots in Figures 13.1 and 13.2 that fifth-grade students located in a textbook (Eicholz, 1991) while searching for graphs they could use to compare two data sets (Russell et al., 2003c). Both are described as "double bar graphs." On quick inspection, they seem to represent data in exactly the same way. However, a closer inspection reveals an important difference. Figure 13.1 uses a bar to represent the cost of each individual item (e.g., hat, shoes). The length of each bar is proportional to that item's cost. We prefer to call this representation a *case-value* plot. This term refers to the fact that each case (e.g., a hat, a cape) is represented by a separate bar indicating that case's value (i.e., price in dollars). Thus the bar's length shows the magnitude or value for that case—a "case value."

The graph in Figure 13.2 shows the number of respondents who chose various personalities as people they would most like to have known. In this type of graph, called a *frequency* bar graph, a bar's height is not the value of an individual case but rather the number (frequency) of cases (respondents) that all have a particular value. In this instance, the leftmost bar shows that 16 respondents (cases) selected Susan B. Anthony. Of course, we could use a stack of Xs instead of bars to represent the number who selected each famous person. Some refer to plots made with Xs as *line* plots. We prefer to call them stacked dot plots.

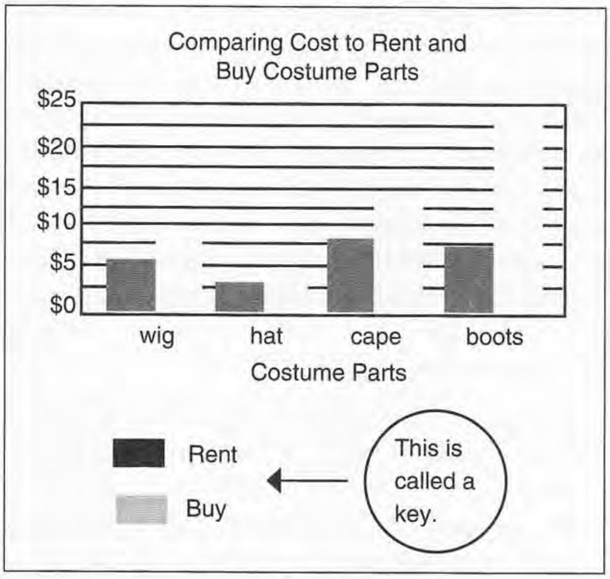

Figure 13.1. A case-value plot of the cost of renting versus buying various costume parts; from Eicholz (1991, p. 96).

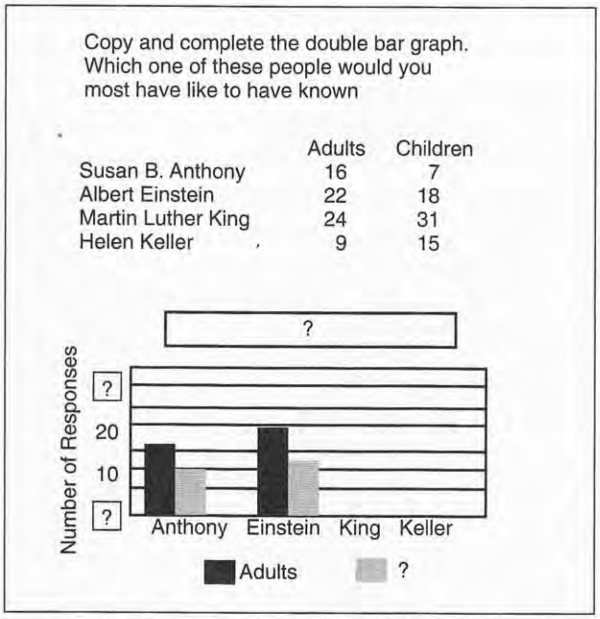

Figure 13.2. A frequency bar graph of the number of adults and children who named various personalities as people they would like to have known; adapted from Eicholz (1991, p. 97).

Anticipating, as Bright and Friel (1998) reported, that students would have more difficulty working with frequency bar graphs and stacked dot plots than with case-value plots, Cobb and his associates had students in their seventh-grade teaching experiment spend the first few weeks working solely with case-value plots (Cobb, 1999). The materials and software "mini" tool they used displayed these bars horizontally to facilitate the later transition to stacked dot plots. Furthermore, they started with problems in which it was relatively natural to display individual data values as length (e.g., length of lizards, braking distance of cars, hours of service from a battery).

After a few weeks working with case-value plots, the researchers introduced students to the stacked dot plot by demonstrating how it could be created from a case-value plot. This introduction involved a progression of intermediate stages: first the bars were erased and only their endpoints plotted; and finally, these "floating" points were allowed to collapse down onto the horizontal axis such that cases with the same values would stack on top of one another.

We demonstrate this progression using weight (in pounds) of the backpacks carried by 55 elementary school students. Figure 13.3 is a case-value plot of the data, with the cases ordered from lightest backpacks (bottom) to the heaviest (top).

In Figure 13.4, each bar in Figure 13.3 has been replaced by a small circle located at what was the right-hand end of the bar. The value of each case is indicated by a circle's position over the x-axis.

Finally, in Figure 13.5, the circles of Figure 13.4 have collapsed onto the x-axis to form stacks showing the number of students with backpacks of a given weight.

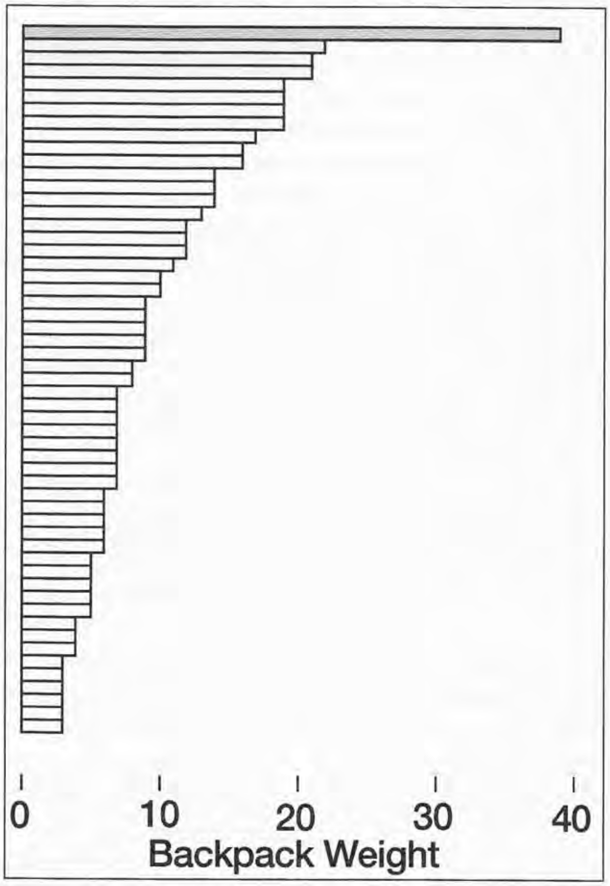

Figure 13.3. Case-value plots showing weight (in pounds) of the backpacks of 55 elementary school children, plotted with Tinkerplots (Konold & Miller, 2002).

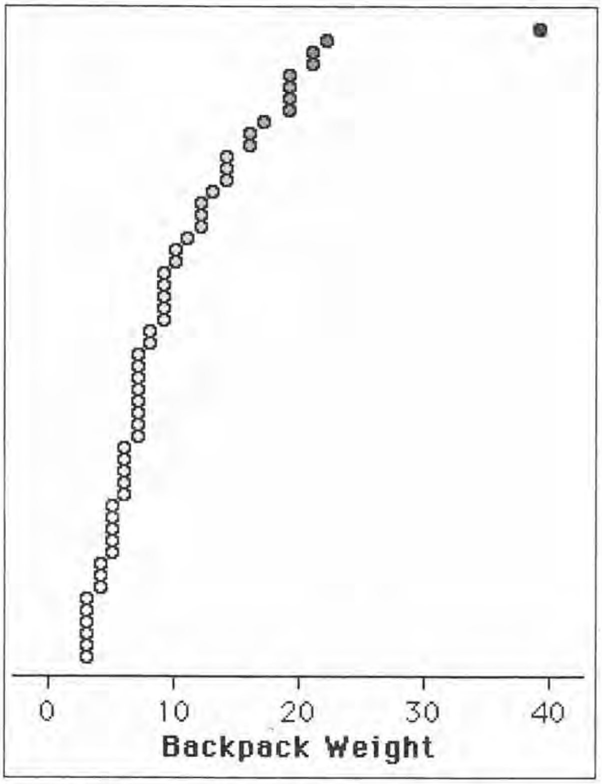

Figure 13.4. Intermediate step in the animation of the case-value plot of Figure 13.3 to the frequency graph of Figure 13.5.

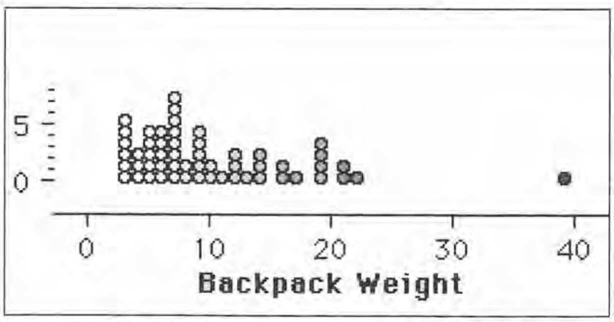

Figure 13.5. Final step in the animation of the case-value plot in Figure 13.3 to a stacked dot plot.

Cobb and his associates believe that by grounding the stacked dot plot in the case-value plot, students were able to view the Xs (or circles) in stacked dot plots as magnitudes (Cobb, 1999). As we discuss next, many of the difficulties students have in representing and interpreting data with stacked dot plots (e.g., putting a value of 12 next to 10 when there is no value of 11) stem from their failure to view the Xs as measures of magnitude.

Deciding About Scale

Deciding on plot scales and on what data should be included in their plots poses a number of interesting challenges to students. Some students described in Russell et al. (2002a) thought that plot scales should not extend beyond the range of observed values, whereas others argued that the scales should extend to include values that could have occurred or far enough at least to make a pleasant boundary.

Working with categorical data, many students argue that they should not include categories with frequencies of zero. Frequencies of zero also raise important scaling issues when students are working with numerical data. Students may be unsure of how or whether to plot nonoccurring data values. For example, some fourth graders (Russell et al., 2002a) used the scale "50, 52, 54, 55, 56, 57, 58, 59, 60, 62, 64" to represent data about classmates' heights on a stacked dot plot, omitting from the axis any value they had not observed. Without gaps along the scale representing nonoccurring heights, it is difficult to see clumping in the data or to judge the magnitude of the difference between various heights. Some of the difficulties students have may result from the fact that they often construct their scales at the same time they plot the data. This practice may make it harder for them to distinguish between the numbers along their scale and the actual data they are plotting.

In making graphs, some students are driven more by conventions they have learned than by the information they want to convey. As Roth and McGinn (1997) pointed out, "In schools...students make graphs for the purpose of making graphs" (p. 95). Students are well practiced, therefore, in setting aside their own intentions and purposes and getting down to the business of producing "good" graphs. This orientation is evident in much of the dialogue among students in the study by Russell et al. (2002a) as they discussed not what they *could* do in plotting their data but what they *ought* to do. Scaling decisions represent ideas about how to frame the data. Thus, they should be tailored to the questions one wants to answer.

Imagine drawing a face on an inflated balloon and then popping it. Next imagine stretching that popped balloon over various objects—a book, a doorknob, a basketball. The face would look quite different stretched over these various objects. The issues in selecting scales for plotting data are exactly the same as those in deciding between doorknobs and basketballs in the balloon example. What will be the minimum and maximum values on the axes, what will be the interval sizes between numbers, and what ratio should we use for the relative sizes of the x- and y-axes (e.g., do we want tall, skinny bars or short, fat ones)? Each of these choices affects how the data appear. In making these decisions, therefore, we cannot afford to adhere to strict conventions but must be mindful of our particular purposes. And we have no reason to puzzle long over these questions. Unlike the balloon example, we have no ideal scale that will make the data appear as they "really" are. We can try out several alternative plots and scales and learn what we can from each. When it comes time to summarize our results for others, we can then pick those representations that do the best job of telling the story sharply and fairly.

Describing and Interpreting Data

David Moore (1990) proposed five core ideas of statistics. Topping his list is the awareness that variation is everywhere. "Individuals are variable; repeated measurements on the same individual are variable" (p. 135). The idea that individuals vary is apparent to even young students. Variability is literally everywhere around them. Their classmates come in a variety of heights, hair colors, and temperaments. The weather they encounter when they step outside varies not only from season to season but from day to day, and sometimes from one minute to the next. If students know nothing else when they begin collecting data, they know that they will get a variety of values.

Anticipating and observing variability within a group are not difficult. What is difficult is figuring out how to quantify variability and perceive and characterize the group as a whole when individuals in that group differ from one another. In this section, we explore how students who tend first to describe and reason only about individual cases begin to describe group characteristics, to summarize and compare groups despite the fact that individuals in each group vary.

Describing Group Features

In early experiences with data, students tend to focus on describing individual data points, or clusters of similar individuals. Russell et al. (2002a, Case 7) described a kindergarten episode in which students reported their favorite color. As the teacher recorded the information on the board, students spontaneously commented on which color was ahead—the modal value. However, the next day when the teacher asked what the graph showed, none of the students mentioned the mode. Instead they made such comments as these:

- We know what everyone's favorite color is.
- My favorite color is red.
- My shirt is blue.
- We learned English and Chinese colors. (Case 7)

Here again we see students not distinguishing all the things they knew about colors and color preferences from the specific data they had collected. Additionally, they focused almost exclusively on individual values. The teacher wondered why it was so obvious to her that blue was the favorite color and why her students "did not seem to pull the individual pieces of information together to share ideas about the data as a whole" (Russell et al., 2002a, Case 7).

However, it is unlikely that the mode these students attended to the previous day as the teacher recorded their responses on the board was, for them, a characteristic or feature of the "data as a whole." Rather, it was simply "the winner." As Mokros and Russell (1995) found in their interviews with students in Grades 4, 6, and 8, those students who used modes to describe data viewed the mode only as the most frequent value. They showed no inclination to use that value to represent or summarize the other values as well.

Students need to make a conceptual leap to move from seeing data as an amalgam of individuals each with its own characteristics to seeing the data as an aggregate, a group with emergent properties that often are not evident in any individual member. "To be able to think about the aggregate," claimed Hancock and colleagues (1992), "the aggregate must be 'constructed.'" (p. 355).

This leap is a difficult transition. Hancock et al. (1992) reported their experiences working in an after-school setting with small groups of students ages 8 to 15. They involved the students in designing statistical questions, collecting data, and exploring data using the software tool Tabletop, which was then under development. The teaching staff explicitly encouraged the use of distributional terms, such as *cluster* and *range*, to characterize data but reported that despite this emphasis, "students often focused on individual cases and sometimes had difficulty looking beyond the particulars of a single case to a generalized picture of the group" (p. 354).

The researchers characterized this individual-based analysis as resulting in blow-by-blow descriptions of results: "This person said 'yes' to Question 1 and to Question 2, but this person said 'yes' for Question 1 but she didn't say 'yes' for Question 2..." (Hancock et al., 1992, p. 354). We see similar responses throughout the cases in Russell et al. (2002a), especially in the earlier grades. Asked what they learned from their survey about who liked their vacation, one kindergartner responded: "11 people said yes, 2 people said no, and 8 people said something else. That makes 21" (Case 9). As a kindergarten teacher observes, this individualistic orientation seems to preclude students from seeing group features:

> In the end, they seemed to attend to names on the chart and the information that was recorded about each person. They did not seem to pull the individual pieces of information together to share ideas about the data as a whole. (Russell et al., 2002a, Case 7)

Hancock and his colleagues (1992) reported similar observations. For example, Tabletop allowed students to display labels of individual data points. When they could, students kept a label showing the identity of each data point (e.g., the names of individual respondents). Worried that students were focusing primarily on individual data points, the teaching staff encouraged students to remove these labels, hoping that doing so would make the features of the whole distribution more salient. However, students could usually remember which data point belonged to which individual and therefore continued to draw on this information in interpreting the data.

With the individuals as the foci, it is difficult to see the forest for the trees. If the data values students are considering vary, however, why should they regard or think about those values as a whole? Furthermore, the answers to many of the questions that interested students—for instance, Who is tallest? Who has the most? Who else is like me?—require locating individuals, especially themselves, within the group. We should not expect students to begin focusing on group characteristics until they have a reason to do so, until they have a question whose answer requires describing features of the distribution. As we discuss subsequently, asking if two groups differ may be one such question.

Types of averages. Only one type of average—the mode—is useful in describing qualitative data, such as favorite color. With numeric data, on the other hand, there are many ways we can characterize the center of distributions. Statistics instruction has traditionally focused on the mean (arithmetic average), median, and mode, with special emphasis on the mean. There are, in fact, many other types of averages. The geometric mean, for example, is often used in economics to find average growth rates and is computed by multiplying all n values and then finding the nth root of the product.

By second or third grade most children have heard of grade averages and average temperatures. Although familiar with the term, their ideas about average are based on everyday meanings that draw on qualitative, rather than quantitative, notions of typicality. In the study by Russell et al. (2002a), we see students using averages of all sorts—means, medians, and especially modes. In addition, we see some using the midpoint between two extreme values (the midrange) and various other intuitive notions, such as the "modal clump" (Konold et al., 2002).

The ideal average that young students appear to have in mind is an actual value in the data set that is also the most frequently occurring value (the mode), positioned midway between the two extremes both in terms of value (the midrange) and order (the median), and that is not too far from all the other cases. In symmetric, mound-shaped distributions with lots of data, one can often find an average value that has all, or nearly all, these properties. But with many of the small data sets students explore, most of these conditions cannot be met. And when students have to start giving up items on their "average" wish list, most of them hang tenaciously onto the mode. Thus, teachers in the Russell et al. (2002a) study describe their students as heading "straight for the mode" and considering it "the end-all way to describe what's typical in a set of data" (Case 12, Case 19).

Mokros and Russell (1995) found in their interviews that most of the fourth graders, and even a few older students at Grades 6 and 8, used the mode in situations in which its use was inappropriate. For example, they asked children to price nine bags of potato chips such that the "typical or usual or average" price of the nine bags would be \$1.38 without any of the individual prices' being \$1.38. Students who conceived of averages only in terms of the most common value were unable to solve the task without using \$1.38 as one of the nine prices.

The mean is noticeably absent from the foregoing list of students' ideal average. We get some insight as to why from the research of Strauss and Bichler (1988). As part of their study of student understanding of the mean, they described seven fundamental properties of the mean. Of those properties, the only one that is clearly among those students regard as important is that the mean is located between (though not necessarily midway between) the extreme values. Two of the seven properties—(1) that the mean does not necessarily equal one of the values in the data set and (2) that it does not necessarily have any counterpart in physical reality—are, in fact, two of the reasons students give for dismissing the mean as a useful average. For example, in Case 25 of Russell et al. (2002a), a third grader objected to using the mean of his two attempts at blowing a Styrofoam cylinder as far as he could because "I didn't get that as one of my distances. It wouldn't be true. It's a lie!"

Traditionally, students in U.S. schools have been introduced to the mean beginning at Grades 4 or 5. In their survey of 122 children in Philadelphia schools, Gal, Rothschild, and Wagner (1990) found that 2% of the third graders, 85% of the sixth graders, and 90% of the ninth graders referred to the add-and-divide algorithm when asked how weather forecasters determined average temperatures. Most researchers who have explored students' use and understanding of means are now recommending that we place much less emphasis in the elementary grades on teaching the mean. As Mokros and Russell (1995) found, those students who do know the standard algorithm for computing the mean generally misunderstand and misapply it. They speculated that "premature introduction of the algorithm ... may cause a short circuit in the reasoning of some children" (p. 37).

Additional research involving students from middle school through college suggests that few students know much more about means than how to compute them (Cai, 1998; Pollatsek, Lima, & Well, 1987). Thus, even children who can compute the mean often do not understand why the computation works or what the result represents. Although the add-and-divide algorithm is relatively simple to execute, developing a conceptual underpinning that allows one to use the mean sensibly is surprisingly difficult. Based on observations of teachers in the study by Russell et al. (2002a), the same difficulties appear in trying to teach students to use the median.

Students frequently use the midrange as an average. Although it may seem a crude average at best, it can be a useful index with large data sets. Eisenhart (1971) pointed out that long before astronomers used means of multiple observations to estimate the actual position of stars, they used the midrange as their estimate. This and other evidence suggests that, historically speaking, the midrange was the precursor to the mean.

A type of average that Russell et al. (2002a) observed many students using is what one third grader called the "middle clump" (Case 21). A middle clump is a cluster of values in the heart of the distribution. For example, a fourth-grade teacher had her students make a stacked dot plot of the number of people in their families. Using stars for emphasis, the students wrote the following statements below their plot (Russell et al., 2002a, Case 3):

- One person has 18 in her family.
- The range of the data 4–18.
- Most typical number of people in the family is 5 or 6.

This bullet summary includes descriptions of spread, center, and a value of special interest.

Additional research by Konold et al. (2002) suggested that students often select a "modal clump" so that it includes all, or most, of the ideal features of averages listed above. The clump of 5 to 6 in the distribution of family sizes in Russell et al. (2002a, Case 3) included the mode and the median and was near most of the data: two thirds of the cases were in the interval 4 to 7. In describing a distribution, statisticians often specify values for both center and spread. They might summarize this distribution by saying that its median is 6 and that the middle 50% of the data (the interquartile range) lies between 5 and 9. Konold et al. (2002) suggested that the modal clump potentially serves a somewhat similar purpose for students, letting them express at the same time both what is average and how variable the data are.

For example, in Case 12 of Russell et al. (2002a), third- and fourth-grade students worked with data indicating the number of years students' families had lived in town (see Figure 13.6).

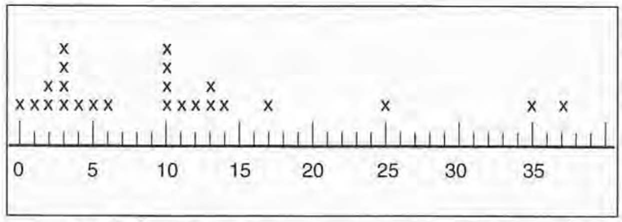

Figure 13.6. Number of years students' families had lived in town.

Excerpts from group reports and class discussion suggest that students were using the idea of modal clumps to make sense of the data as a whole:

Zia & Tuyet: Not a lot of people have lived here for a long time. A lot of people have been here for three years. There are two big clumps and a lot of gaps.

Anna, Amber, & Janis: At the beginning we can plainly see that there is a very big clump of 0, 1, 2, 3, 4, and 5.

Kevin: Most of the Xs are between 0 and 6. It's the biggest clump.

Anna: There's eleven [in the clump]. That's almost half.

Reflecting on how her students had summarized the data, the teacher noted:

> That first big clump clearly needed to be part of it [the summary] in the kids' eyes, and the fact that it also contained that mode at 3 didn't hurt either.... "Almost half" was good enough for them. It seemed to carry some weight of significance, and as I thought about it, I realized that it did for me also. This was a meaningful statement to make about our data.... (Russell et al., 2002a, Case 12).

Cobb (1999) described how the seventh graders in his teaching experiment began reasoning about data sets as wholes once they were able to isolate and converse about what they called the "hills" in the data. Many researchers have emphasized that we should encourage students to use less formal methods like these, and that in many situations what students come up with as descriptors of average are as good as, or better than, the mean or median in summarizing the data (Bakker, 2001; Cobb, 1999; Mokros & Russell, 1995). Not only are such averages as modal clumps (or hills) often good enough for the task at hand, but such ideas can provide the bases for later constructing meaningful interpretations of more traditional averages, such as means and medians.

Interpretations of averages. Of course, as with any average, not all students who describe a modal clump are thinking of it as a description of the group as a whole. Just as they often do with the mode, some students think of the modal clump simply as the "winner." Indeed, they may be using the clump as an extended mode, as a description of the most common values.

To explore further how students think about and use averages, we need to distinguish more carefully between the types of averages they use (modes, medians, midranges, modal clumps, etc.) and the meanings they give those averages. A fourth-grade teacher in the study by Russell et al. (2002a, Case 26) made this distinction as she tried to explore what significance her students gave to the mean they had just computed. After Robert answered that they had "just used a calculator" to get the average, the teacher made this clarification:

Teacher: I'm actually not asking you *how* you got the answer. I'm asking what your answer means. Do you have a sense of that?

Robert: Not really.

We might argue that Robert interpreted the mean simply as a result of a computation, which would explain why he answered the way he did. Earlier in the episode, others had questioned why Robert and his group had used only four values to get an average to decide how many fourth graders they would need to build a 100-foot tidal wave. To others in the class, this sample seemed too small. Robert replied, "You don't need a whole lot of people to get an average," by which he may have meant that to get an answer by adding and dividing, four will do. Of the 20 students Mokros and Russell (1995) interviewed, 3 seemed to view averages just as Robert did, simply as a procedure. The researchers described this approach as "algorithmic" and described it along with four other approaches that they observed students using.

Konold and Pollatsek (2002) have suggested a somewhat different set of "interpretations," which include among them ones that more experienced data analysts employ. Konold and Pollatsek (2002) characterized an interpretation as "the goal a person has in mind when he or she computes or uses an average. It is the answer a person might give to the question 'Why did you compute the average of those values?'" (p. 267). The interpretations they describe include average as a data reducer, as a fair share, and as a typical score.

Viewed as a data reducer, an average is a way to boil down a set of numbers into one value. A high school student interviewed by Konold, Pollatsek, Well, and Gagnon (1997) gave the following rationale for why she would use a mean or median to describe the number of hours worked by students at her school:

> We could look at the mean of the hours they worked, or the median... It would go through a lot to see what every, each person works. I mean, that's kind of a lot, but you could look at the mean. (Konold and Pollatsek, 2002, p. 268)

The data need to be reduced because of their complexity, in particular because of the difficulty of holding in memory the individual values. The student quoted above went on to say, "You could just go through every one... [but] you're not going to remember all that."

An average is interpreted as a fair share when we imagine redistributing a quantity among individuals so that in the end each has the same amount. The computation for the mean is probably first encountered in elementary school in the context of fair-share problems, with no reference to the result's being a mean or average. Most of the tasks Strauss and Bichler (1988) used in their research were problems that seemed to suggest a fair-share interpretation, such as in the following example:

> Ruth brought 5 pieces of candy, Yael brought 10 pieces, Nadav brought 20, and Ami brought 25. The children who brought many gave some to those who brought few until everyone had the same number of candies. How many candies did each girl end up with? (Adapted from Strauss & Bichler, 1988)

An average interpreted as a typical score, Konold and Pollatsek (2002) suggested, includes ideas related to the majority, mode, median, and midrange. Teachers in the Russell et al. (2002a) casebook often posed questions to students hoping to elicit this interpretation of a "typical" or "representative" value. The teacher posing a question such as "How tall is a typical fourth grader?" is presumably thinking of a value that is representative of the entire group.

However, many of the students' responses suggested they believed a typical value described a characteristic of a particular case, or set of cases, in the distribution. A third-grade teacher (Russell et al., 2002a, Case 22) asked her class, "What would you say is the average height of kids in our room?" Brita volunteered, "It's me. I think I am average." She seemed not to be focusing on an average height but rather on a characteristic of a person: "I'm average." After students lined up by height, other students used a similar notion in describing average, viewing average as a characteristic of a person rather than the group.

> *Phoebe:* I think I'm taller than average because I notice that on the playground.
>
> *Brita:* I was right. Sam is average, and I'm average too. We are the same.
>
> *Tiffany:* I'm average too.
>
> *Katie:* I'm not average. I'm shorter.

To claim that Sam is of "typical" or "average" height is to characterize him and not necessarily the group as a whole. We are unsure whether these students would consider Sam's height to be a good characterization of the whole group any more than they would consider Katie's diminutive height to be a useful summary. This use of averages to describe particular individuals rather than the group is supported by common usage in which we frequently speak of the "average" or "typical" student.

Mokros and Russell (1995) suggested that students who are beginning to make sense of averages as representative values may draw on intuitive ideas about what seems "reasonable." Students using this "reasonable" approach drew on their experiences with the phenomenon to explain, for example, why all the values would not be the same. This approach was used by 20% of the students they interviewed. Here is how a fourth grader described the distribution of allowances she had created by placing individual tiles on a stacked dot plot such that the average would be $1.50:

> [It depends] on how old they are.... If there are some kids that were like 15 and 16 years old and there are other kids that were 10 years old.... It depends on how rich their parents are sometimes.... If the typical [allowance] is $1.50, you're not going to really go above $5.00 for any kid. If I got $5.00, it would be good.... And you know that when you run around with a lot of kids, most of them are like $1.50 or $1.75 or $1.25 or $1.00, something like that. (p. 30)

Students using this approach also relied on intuitions that averages are roughly in the center, thinking of an average not as a precise location but as an around-about sort of thing.

> *Interviewer:* Tell me how you're thinking about this one.
>
> *Suzanne:* Well, just as they get higher, sometimes they should get lower. And you said the typical allowance is about $1.50, so some kids can get $1.50. And if it were $1.75 that would be pretty close and so would [$1.25], because that's around it.... Parents don't like to waste their money on kids. (p. 30)

Comparing Groups

If we consider statistics instruction as a staircase beginning in early elementary school and continuing up to Grade 12, the ability to compare two groups should be seen as a major landing midway up the stairs. We might think of it as the place where instruction in the early years is headed and as the foundation from which further statistics will rise. Making such comparisons is the heart of statistics. Most of the important issues and questions argued with data amount to comparing two groups, for example, treatment and control groups in medicine, before-and-after groups in various interventions and educational studies, and females versus males in gender equity studies. Furthermore, students will not understand the rationale of statistical inference until they first are comfortable summarizing a difference by comparing two groups using some measure of center. Nor are the skills required in comparing groups separate from those that allow us to perceive trends and patterns in the first place. When we go snooping about in data, we are searching for just these sorts of group differences, and the question is, How do we spot them?

Research has demonstrated that students do not initially know how to approach a group-comparison problem. This is true even of students who appear to know quite a bit about averages. Gal, Rothschild and Wagner (1990) interviewed students in Grades 3, 6, and 9 to determine, among other things, their understanding of how means were computed and used. They also gave the students nine pairs of stacked dot plots. In one version of their materials, graphs showed the results of a frog-leaping contest between two teams, with Xs on the graphs representing the distances jumped by individual frogs of each team. The students' task was to decide from the data which team won the contest. Only half of the sixth- and ninth-grade students who knew how to compute means went on to use means to compare the two groups, even when the groups were of unequal size.

A number of studies have observed students who appeared to use averages to describe a single group but did not use them to compare two groups (Hancock et al., 1992; Watson & Moritz, 1999; Jones et al., 1999; Konold et al., 2002). Bright and Friel (1998), for example, questioned eighth-grade students about a stem-and-leaf plot that showed the heights of 28 students who did not play basketball. They then showed them a stem-and-leaf plot that included these data along with the heights of 23 basketball players. The heights of basketball players were indicated, as they are in Figure 13.7, in bold type.

10	
11	
12	
13	8 8 8 9
14	1 2 4 7 7 7
15	0 0 1 1 1 1 2 2 2 2 3 3 5 6 6 7 8
16	
17	1
18	**0 3 5**
19	**0 2 5 7 8 8**
20	**0 0 2 3 5 5 5 5 7**
21	**0 0 0 5**
22	**0**
23	

Figure 13.7. Heights of students and basketball players (bold); adapted from Bright and Friel (1998, p. 81). The row headed by 13 (the stem) contains four cases (leaves), three students each of 138 centimeters, and a fourth student of 139 centimeters.

Asked about the "typical height" in the distribution of the heights of nonplayers, two of four interviewed students specified a modal clump (e.g., 147–151 cm). But shown the plot with both distributions, these students could not generalize their method to determine "How much taller are basketball players than students?" The students who did make comparisons compared selected individuals from each group (e.g., pointed out that the tallest student was shorter than the shortest basketball player). In the words of Bright and Friel (1998), some of these students could "describe a 'typical' student or basketball player, but they did not make the inference that the 'typical difference' in heights could be represented by the 'difference in typicals'" (p. 80).

Konold et al. (1997) found similar difficulties among high school seniors who had just completed a yearlong course in probability and statistics. On many occasions during the course, the students had used both medians and means to compare groups. But in the classroom they were supported by the curricula, software, and instructor and thus to a large extent were not choosing for themselves the methods they would use. During a postcourse interview, where they were less constrained in their choice of methods, they seldom used medians, means, or percentages when comparing two groups, though they did use them when summarizing single groups.

Using averages to compare two groups requires viewing averages as a way to represent or describe the entire group and not just a part of it. For this reason, Konold et al. (1997) argued that students' reluctance to use averages to compare two groups suggests that they have not developed a sense of average as a measure of a group characteristic that can be used to represent the group. Thus, students may use averages as part of a description of a single group and yet not accept them as representative of that group. The students interviewed by Bright and Friel (1998) may have been thinking of typicality as a characteristic of the heights of just those people in the center of a distribution. If so, then it is understandable that they would not consider using that modal clump as a characteristic or measure of the whole group for the purpose of comparing it to another group.

The cases in Russell et al. (2002a) demonstrate the power of comparison tasks to engage student interest. Even more important, they appear to prompt students to shift their focus from individuals in a data set to the group as a whole. However, as a teacher notes in Case 27, comparison problems can be "a little puzzling to kids at this age [Grades 3 and 4]—how can you talk about the group, after all, as something somewhat separate from the individuals in the group?" Indeed, the power of comparison questions is that they provide a clear problem that must be resolved—the problem of what to use as group measures that can be compared.

The teaching experiment by Cobb (1999) does suggest that students can appropriate the idea of modal clumps for the purpose of comparing groups. As the researchers were designing the teaching experiment, they had planned on developing the median as the primary indicator of center. However, though the median was a frequent topic of class discussion, students rarely used it to compare groups. Students made most of their decisions about group difference by comparing the numbers of individuals in each group within narrow slices of the range. Konold et al. (1997) reported the same tendency to compare slices among high school students , even when the two groups were of radically different sizes.

Cobb (1999) described a critical episode during the seventh-grade teaching experiment when a student, Janice, used the idea of modal clumps, or "hills" as she called them, to compare two groups. The class was investigating speeds of two groups of cars sampled before and after a police speed trap. During a class group discussion, Janice suggested:

> If you look at the graphs and look at them like hills, then for the before group, the speeds are spread out and more than 55, and if you look at the after graph, then more people are bunched up close to the speed limit [50 mph], which means that the majority of the people slowed down close to the speed limit. (p. 19)

Cobb (1999) reported that this was the first time in a whole-group discussion that a student had "described a data set in global, qualitative terms by referring to its shape" (p. 19). This language was soon taken up by other students in the class and became a standard way for them to describe and compare groups. As they progressed to comparing data sets of different sizes, they began talking not just about the location of two hills but about the number of cases in a hill relative to the number of cases in its group.

Judgments About Covariation

In analyzing data, we frequently want to know whether and how two things are related. Are increases in air-borne pollutants related to warmer-than-average temperatures? Does the death penalty act as a deterrent to serious crime? Is classroom size related to student achievement? In the previous section, we focused on group-comparison questions in which one variable was numeric (such as speed of car) and the other was categorical (such as timing of observation). We can think about comparing such groups as trying to decide whether two variables are related—whether their values covary. Concluding from data that cars on a certain stretch of road tend to go slower after a speed trap than before implies that a car's speed varies with the time of observation (before versus after the speed trap).

In this section, we review research related to judgments about two other types of covariation: determining whether a relationship exists between (1) two categorical variables (e.g., Is having a curfew related to gender?) and (2) two numeric variables (e.g., Does blood pressure increase with age?). As in the previous section, we concern ourselves here with how students learn to examine data in an attempt to detect such relationships. We do not review the research on how students decide whether a trend is strong enough (i.e., statistically significant) to justify concluding that the relationship exists.

Comparing two categorical variables. To make judgments about the relation between two categorical variables, we often display data in a contingency table. Instructors of introductory college courses are often heard to say that contingency tables are the hardest displays their students encounter. The difficulty of reasoning about such data has been borne out in numerous studies.

For example, Batanero, Estepa, Godino and Green (1996) gave data like those in Table 13.1 to more than 200 Spanish students in their final year of secondary school. Their task was to decide whether the data indicated a relationship between the two variables. Most of the students had had instruction in elementary statistics and probability, but they had not yet been introduced to contingency tables.

TABLE 13.1. Incidents of bronchial disease among smokers and nonsmokers; adapted from Batanero et al. (1996).

	Bronchial disease	No bronchial disease	Total
Smoke	90	60	150
Not smoke	60	40	100
Total	150	100	250

According to the data in Table 13.1, smoking is not associated with bronchial disease. One way to see this lack of dependency is by comparing the percentage of the two groups who have the disease. Of the 150 smokers, 90 (or 60%) have bronchial disease, but 60% of the nonsmokers also have it. We would therefore conclude on the basis of these data that smoking has no apparent effect on bronchitis (but, of course, these data are fictional).

Students in the Batanero et al. (1996) study solved about 30% of such problems correctly, a high success rate compared to results from similar studies. The majority of incorrect responses were, however, well predicted by previous research. Many students (13% in the Batanero et al. study) seemed to attend only to the number 90 in deciding that smoking is associated with bronchitis. Even more common was for students to attend to the fact that 90 of the 150 smokers had bronchial disease. Because more of the smokers had the disease than did not, many students believed that the data from the smokers alone was sufficient to establish a connection between smoking and the disease.

The high school seniors interviewed after a statistics course by Konold et al. (1997) used a similar strategy in evaluating data collected from students at their school. Two students, M and J, who were analyzing these data using statistical software, posed the question of whether more female students held jobs than male students. They generated the contingency table shown in Table 13.2. The interview excerpt below suggests that these students attended only to the information about those who held jobs and that they viewed the data about students who did not hold jobs as irrelevant to their question.

TABLE 13.2. Incidents of jobs among male and female high school students; adapted from Konold et al. (1997).

	Job		
Sex	No	Yes	Total
F	23 (0.29)	57 (0.71)	80
M	16 (0.22)	57 (0.78)	73
Total	39 (0.25)	114 (0.75)	153

M: So, oh no, it's pretty even.

I: So tell me how—So, you're, you're looking at the percentage of males and—

M: Yeah.

J: Yeah, the difference.

M: Yeah, of females and males who have jobs.

J: Or who don't.

M: And that don't. And for the amount of—that do have jobs, that females and males are pretty even.

I: So what—tell me the numbers. I can't read them.

M: 57 males and 57 females. (p. 158)

In correcting J's *or* to *and*, M seemed to be stressing that the question about males and females who hold jobs was, in her mind, separate from the question about those would did not hold jobs. In her subsequent statements, M made it clear that she was looking at those who held jobs in forming her conclusion. But because the numbers of males and females are different, we cannot simply compare the numbers of males and females who have jobs. We must compare the rate of jobs among males to that among females to answer the question.

We frequently see this kind of reasoning in news reports or advertising where we are given only half the information we need. "Seventy percent of headache suffers who take AcheAway feel relief within an hour." The implication is that AcheAway did the trick. But what percent of suffers who take nothing feel relief within an hour? For all we know, it is 80%, and AcheAway tends to make things worse.

Follow-up research by Batanero, Estepa, and Godino (1997) suggests just how difficult it is for students to learn to reason about relationships between two categorical variables. They worked with 19 high school students for approximately 30 hours in an intervention that included instruction on group differences and detection of covariation in numeric and categorical data. Before instruction, the students could correctly solve about 20% of the problems involving contingency tables. After instruction, this value went up to only about 30%. In contrast, before instruction, students could successfully evaluate relationships in scatterplots 35% of the time; this value increased to 60% after instruction.

Comparing two numeric variables

To explore relationships between two numeric variables, data are traditionally displayed in two dimensions, with each case represented by a point. The horizontal position of a case is determined from its value on one numeric variable, and its vertical position is determined by its value on the other numeric variable. Data that vary over time (time series) are often displayed this way. Because time series are true functions—only one *y* value is associated with every *x* value—we frequently connect the points with a line, which can help us see general trends in the data.

Figure 13.8 shows the time series of gold-medal times for the men's 100-meter dash for all the years since the Olympics began. Ben-Zvi and Arcavi (2001) gave these data to 13-year-old Israeli students who were participating in a teaching experiment. The students' first task was a general one—to learn what they could from these data. Though this plot could be rescaled to make it easier to perceive, we can still detect an overall downward trend, with running times generally improving (i.e., getting smaller) over the years. To see this trend, we must focus on the data as a whole. As we saw with the task of group comparisons, many students attend to particular data values, or pairs of values, and therefore do not perceive such trends.

The students in this teaching experiment, A and D, were working with this plot and a corresponding table of values. They initially focused only on single values.

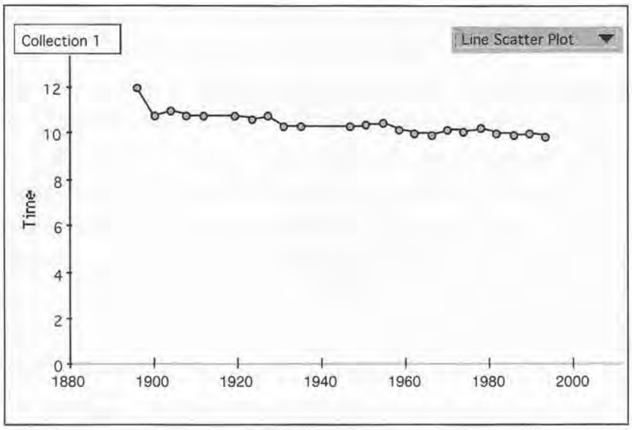

Figure 13.8. Olympic gold medal times for the men's 100-meter dash. Reproduced using Fathom; adapted from Figure 13.2 in Ben-Zvi and Arcavi (2001).

A: [Reading the question] What do you learn from this graph? We learn from the graph, in which year there was which running time.

D: What running time was achieved in what year.

Their teacher intervened in an effort to help them move from these localized perceptions to more global ones.

T: Does it decrease all the time?

A&D: No.

T: No. Does it increase all the time?

A&D: No.

T: No. So, what does it do after all? …

D: It generally changes from Olympiad to Olympiad. Generally, not always. (p. 53)

With additional prompting from their teacher, the students began to notice also that the differences between adjacent years were not constant. The researchers speculated that the students may have been drawing on earlier experience they had had using spreadsheets to explore linear functions in which the differences between adjacent values were constant. This experience may have provided the background against which the fluctuating differences in the Olympic data stood out and became the focal point for the students.

With continued prompting, the students began to take note of the overall trend, concluding in their written report,

> The overall direction is increase in the records, yet there were occasionally lower (slower) results, than the ones achieved in previous Olympiads. (p. 54)

Scatterplot displays are similar to the time-series plot except that for every value of x, multiple values of y may occur. Thus, with scatterplots, not only is the trend not constant, but a scattering of values further masks that trend. The scatterplot in Figure 13.9 shows the relationship between students' weights and the weights of their backpacks for 55 elementary school students. Heavier students tend to carry heavier backpacks, though we can see plenty of exceptions to this trend. This, by the way, is a good example of the importance of the warning that "correlation does not prove causation." It is probably not the case that students carry more weight in their backpacks because they are heavier. Rather, the students are getting heavier as they age and move into the higher grades. And as any sixth grader will tell you, they tend to get more homework assignments than do first graders.

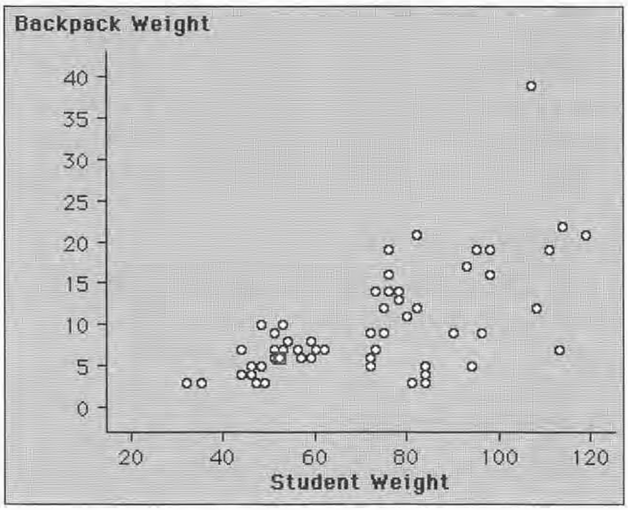

Figure 13.9. The weights (in pounds) of students and their backpacks, plotted using Tinkerplots.

In research on reasoning in the workplace, Noss, Pozzi, and Hoyles (1999) studied the statistical reasoning of practicing nurses. As part of a follow-up teaching experiment, nurses analyzed a health database of British adults. Using statistical software, the nurses quickly generated a scatterplot to answer the question of whether a relationship existed between age and blood pressure. The nurses knew from experience that with increasing age, blood pressure tends to rise. However, they were unable to see this relationship in the scatterplot, the trend apparently masked by the variability in the data. The researchers prompted the nurses to make several vertical slices in the scatterplot, essentially recomposing the continuous age variable into several discrete categories (e.g., ages 18–29, 30–44, 45–59, 60–75). By computing and comparing the average blood pressure in each of these groupings, the nurses then could perceive the trend they had expected.

Cobb and his associates have used this same "slicing" technique in their teaching experiment with middle school students, providing them computer "minitools" that allow students to form such vertical partitions easily (Cobb, McClain, & Gravemeijer, in press). Their intention is to build systematically on understandings that students first develop from comparing two sets of numeric data. By seeing each vertical slice of data in a scatterplot as a distribution of a discrete group, their hope is that students can visually locate the centers (or "hills") of each slice to detect a general trend.

However, scatterplots may not be the best representation for novices to begin using as they explore covariation in two numeric variables. Emerging evidence suggests that it may be easier for students to perceive relationships in two-column tables or case-value plots in which the values of one variable have been ordered and displayed next to the corresponding values of the other variable (see Konold, 2002). Figure 13.10 shows such a paired case-value plot for the same backpack data displayed previously. Knowing that the cases have been ordered according to student weight, one can see the increasing trend of backpack weight in the plot on the right by tracking the changing length of the bars while moving one's eye up the plot.

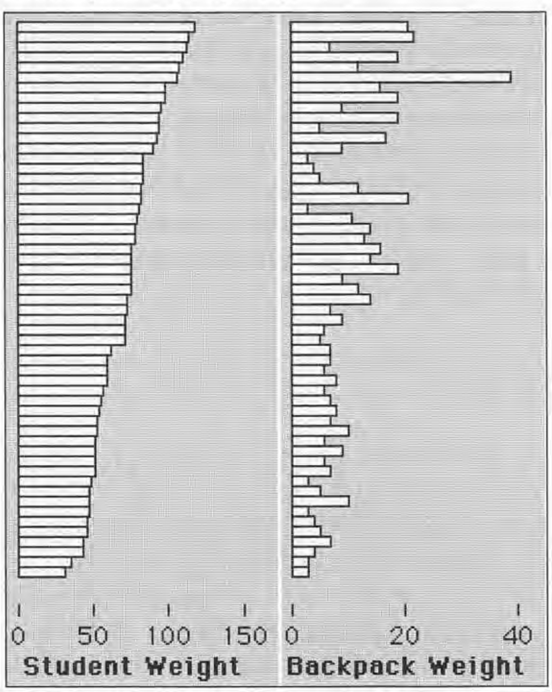

Figure 13.10. The weights (in pounds) of students and their backpacks, plotted using Tinkerplots. The two bars immediately across from each other belong to the same student.

Cobb and his colleagues (in press) mention that before they had introduced their students to scatterplots, they gave them data on carbon dioxide produced by cars going various speeds. They asked students to generate a plot that would allow them to make recommendations about highway speed limits. None of the students made scatterplots; nearly half of them drew a paired double-bar graph like the one shown in Figure 13.10. At the end of the eighth-grade teaching experiment, Cobb and colleagues interviewed 11 of the students. The researchers gave students data from a similar context and asked them to choose from among several representations the one that would best help them judge whether there was a relationship between the two variables. Even though the scatterplot was the only representation they had worked with during the 14-week teaching experiment, only four of the students chose a scatterplot display to make that judgment. The other students chose either a paired double-bar graph or the corresponding table of paired values (J. Cortina, personal communication, 21 December 2000).

Relating Data to the Observed Event

We have discussed some of the issues that arise as students generate, represent, and interpret authentic or meaningful data. Previously we emphasized that in making sense of data, students need to develop an understanding of data as being related to, but not identical with, real events—as models of those events. After reviewing some of the complexities students face as they construct the idea of data as an aggregate and begin to work through problems of data analysis, we find that we have come full circle. The further students become immersed in the tools of data analysis, the more important it becomes that they maintain the connection between the data, the events they represent, and the question that motivated their construction. Too often students treat data as numbers only, forgetting that those numbers have a context and that the reason for analyzing them is to learn more about that context. Students seem particularly vulnerable to treating data as numbers only when they work with data they themselves have not collected.

Cobb (1999) reported that before instruction, the seventh-grade students with whom he and his colleagues worked tended to view data analysis as "doing something with the numbers" (p. 12). In summarizing student responses, Cobb concluded that it was "doubtful whether most of the students were actually analyzing data, in that the numbers they manipulated did not appear to signify measures of attributes of a situation about which a decision was to be made" (p. 13). Early in the subsequent teaching experiment, the researchers saw the same tendency as students began reasoning about the data in Figure 13.11, which showed hours of use of "Always Ready" and "Tough Cell" batteries displayed as a case-value plot.

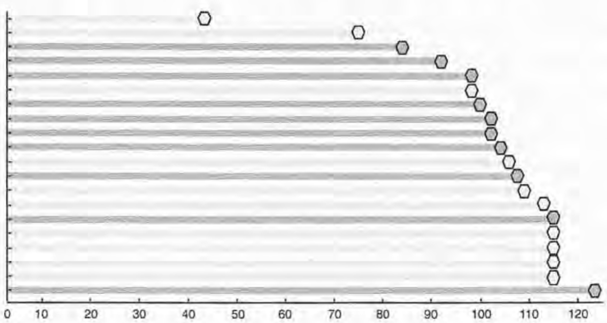

Figure 13.11. Case-value plot of hours of use of "Always Ready" (light gray) and "Tough Cell" (dark gray) batteries; adapted from Cobb (1999).

The two brands of batteries appeared as green and pink bars on the computer screen the students were viewing. In Figure 13.11, as well as in the class dialogue below, we have changed these colors to light and dark gray, respectively. During the first day working with this display, the students referred mostly to numbers and colors. Noticing this tendency, the teacher began to encourage them to talk instead about *batteries*.

Casey: And I was saying, see like there's 7 [light gray] that last longer.

Teacher: OK, the [light gray] are the Always Ready, so let's make sure we keep up with which is which. OK?

Casey: OK, the Always Ready are more consistent with the 7 right there, and then 7 of the Tough ones are like further back, I was just saying 'cause like 7 out of 10 of the [light gray] were the longest, and like…

Ken: Good point.

Janice: I understand.

Teacher: You understand? OK, Janice, I'm not sure I do, so could you say it for me?

Janice: She's saying that out of 10 of the batteries that lasted the longest, 7 of them are [light gray], and that's the most number, so the Always Ready batteries are better because more of those batteries lasted longer. (pp. 14–15)

Cobb (1999) concluded that a an essential step in these students' learning to reason about data was coming to expect that statements and claims about various data displays should extend beyond mere numbers by making reference to a specific real-world situation.

In most of the activities described in the cases compiled by Russell et al. (2002a), students collected their own data. Yet we see clear examples of students' losing the connection between the data and the real-world situation. The failed connection often occurs as students begin learning how to describe general features of the data. For example, in Case 23, third-grade students wrote summaries describing a stacked dot plot of daily temperatures they had collected in February. One student described the data this way: "At first very spread out. Then it gets more bunched up."

Concerning the student who wrote this summary, the teacher wondered,

> Did he know it wasn't just a clump of Xs, but a representation of a real thing, which was indicating a predominance of a certain temperature on the high side of the range of temperatures for the month? … I wondered how to help him see that what he noticed about how the data *looked* implied something significant about what the temperature was like in February. (p. 134)

Feldman et al. (2000) cited several examples from data-intensive science projects of what happens when we give students data about phenomena far removed from their experience and with no clear questions to answer. They concluded that

> nearly every problem associated with…keeping them engaged in analysis ultimately stems from students not making, or losing, the connection between the data they have and a real-world question. This being the case, the solution to most of the problems can be found in focusing on how to make and maintain these connections. (p. 127)

Conclusion

The interdependency of the stages of data analysis places high demands on young students and their teachers. The students are only beginning to learn how to turn observations into data and are not yet aware of how they can probe data to answer questions. Unlike the expert, novices have little relevant experience they can use to plan ahead. Helping students raise questions that interest them and that they can productively pursue is a challenge for the teacher. Left on their own, students are often overwhelmed or wander off track; given too much structure and assistance, they can lose sight of the big picture and their motivation for looking at data.

The challenge for teachers is to "help students keep hold of the big picture as they explore the parts" (Russell et al., 2002a, Case 21). Doing so requires finding ways to manage complexity so that students can think about the questions they are pursuing, focusing not on "features of the graph" per se, but on "its implications as a representation of something real."

The research and cases we have presented call attention to the need for students to work with real data throughout the elementary and middle grades. Understanding data representation and analysis involves many complex issues, from sorting through what different numbers on a graph mean, to choosing appropriate measures to summarize and compare groups, to identifying relationships between variables. Through multiple experiences with a variety of data sets, students begin to develop the tools and concepts they need to use data themselves and to interpret the data they will encounter throughout life.

ACKNOWLEDGMENTS

Portions from this chapter appear in an article titled "Highlights of Related Research," by Clifford Konold and Traci L. Higgins, in *Working With Data*, edited by Susan Jo Russell, Deborah Schifter, and Virginia Bastable, pages 165–201 (Parsippany, N. J.: Dale Seymour Publications, 2002). These portions are included here with permission.

The writing of this chapter was supported in part by the National Science Foundation under Grant Nos. REC-9725228, ESI-9730683, ESI-9731064, and ESI-9818946. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Joan Garfield and John Carter provided comments on an earlier version of the chapter.

REFERENCES

American Statistical Association. (1997). *Report from the American Statistical Associations' Advisory Review Group.* Alexandria, VA: Author. Available: http://www.stat.ncsu.edu/stated/standards.html

Bakker, A. (2001). Historical and didactical phenomenology of the average values. In P. Radelet-de Grave (Ed.), *Proceedings of the Conference on History and Epistemology in Mathematics Education* (Vol. 1, pp. 91–106). Louvain-la-Neuve & Leuven, Belgium: Catholic Universities of Louvain-la-Neuve & Leuven.

Batanero, C., Estepa, A., & Godino, J. D. (1997). Evolution of students' understanding of statistical association in a computer-based teaching environment. In J. B. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics: Proceedings of the 1996 IASE Round Table Conference* (pp. 191–205). Voorburg, The Netherlands: International Statistical Institute.

Batanero, C., Estepa, A., Godino, J. D., & Green, D. R. (1996). Intuitive strategies and preconceptions about association in contingency tables. *Journal for Research in Mathematics Education, 27*, 151–169.

Biehler, R. (1989). Educational perspectives on exploratory data analysis. In R. Morris (Ed.), *Studies in mathematics education, Vol. 7: The teaching of statistics* (pp. 185–201). Paris: UNESCO.

Ben-Zvi, D., & Arcavi, A. (2001). Junior high school students' construction of global views of data and data representations. *Educational Studies in Mathematics, 45*, 35–65.

Bright, G. W., & Friel, S. N. (1998). Graphical representations: Helping students interpret data. In S. P. Lajoie (Ed.), *Reflections on statistics: Learning, teaching, and assessment in grades K–12* (pp. 63–88). Mahwah, NJ: Erlbaum.

Cai, J. (1998). Exploring students' conceptual understanding of the averaging algorithm. *School Science and Mathematics, 98*, 93–98.

Cleveland, W. S. (1993). *Visualizing data*. Summit, NJ: Hobart Press.

Cobb, G. W. (1997). Mere literacy is not enough. In L. A. Steen (Ed.), *Why numbers count: Quantitative literacy for tomorrow's America* (pp. 75–90). New York: College Entrance Examination Board.

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

Cobb, P., McClain, K., & Gravemeijer, K. (in press). Learning about statistical covariation. *Cognition and Instruction.*

Eicholz, R. (1991). *Addison-Wesley Mathematics—Grade 5*. Reading, MA: Addison-Wesley.

Eisenhart, C. (1971). *The development of the concept of the best mean of a set of measurements from antiquity to the present day.* Unpublished notes from the presidential address at the 131st meeting of the American Statistical Association, Fort Collins, Colorado.

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

Friel, S. N., & Bright, G. W. (1998). Teach-Stat: A model for professional development in data analysis and statistics for teachers K–6. In S. P. Lajoie (Ed.), *Reflections on statistics: Learning, teaching, and assessment in grades K–12* (pp. 89–117). Mahwah, NJ: Erlbaum.

Gal, I., Rothschild, K., & Wagner, D. A. (1990). *Statistical concepts and statistical reasoning in school children: Convergence or divergence?* Paper presented at the meeting of the American Educational Research Association, Boston.

Gallup, G. (1972). Opinion polling in a democracy. In J. M. Tanur, F. Mosteller, W. H. Kruskal, R. F. Link, R. S. Pieters, & G. R. Rising (Eds.), *Statistics: A guide to the unknown* (pp. 146–152). San Francisco: Holden-Day.

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education, 19*, 44–63.

Hancock, C., Kaput, J. J., & Goldsmith, L. T. (1992). Authentic inquiry with data: Critical barriers to classroom implementation. *Educational Psychologist 27*, 337–364.

Hand, D. J. (1998). Data mining: Statistics and more? *The American Statistician, 52*, 112–118.

Jacobs, V. R. (1999). How do students think about statistical sampling before instruction? *Mathematics Teaching in the Middle School, 5*(4), 240–246, 263.

Jones, G. A., Thornton, C. A., Langrall, C. W., Mooney, E. S., Perry, B., & Putt, I. J. (1999). *A framework for assessing students' statistical thinking.* Paper presented at the meeting of the Research Presession of the National Council of Teachers of Mathematics, San Francisco.

Konold, C. (2002). Teaching concepts rather than conventions. *New England Journal of Mathematics*, 34(2), 69–81.

Konold, C., Higgins, T., & Russell, S. J. (2000). Developing statistical perspectives in the elementary grades. In M. L. Fernandez (Ed.), *Proceedings of the 22nd Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (p. 329). Tucson, Arizona.

Konold, C., & Miller, C. M. (2002). Tinkerplots [Computer software, in development]. Amherst, MA: University of Massachusetts, SRRI.

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

Konold, C., Pollatsek, A., Well, A., & Gagnon, A. (1997). Students analyzing data: Research of critical barriers. In J. B. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics: 1996 Proceedings of the 1996 IASE Round Table Conference* (pp. 151–167). Voorburg, The Netherlands: International Statistical Institute.

Konold, C., Robinson, A., Khalil, K., Pollatsek, A., Well, A., Wing, R., & Mayr, S. (2002). *Students' use of modal clumps to summarize data.* Paper presented at the Sixth International Conference on Teaching Statistics, Cape Town, South Africa.

Lehrer, R., & Romberg, T. (1996). Exploring children's data modeling. *Cognition and Instruction 14* (1), 69–108.

Metz, K. E. (1999). Why sampling works or why it can't: Ideas of young children engaged in research of their own design. In F. Hitt & M. Santos (Eds.), *Proceedings of the 21st Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 492–498), Morelos, Mexico.

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

Moore, D. S. (1990). Uncertainty. In L. A. Steen (Ed.), *On the shoulders of giants* (pp. 95–137). Washington, DC: National Academy Press.

Moore, D. S. (1992). Teaching statistics as a respectable subject. In F. S. Gordon & S. P. Gordon (Eds.), *Statistics for the twenty-first century* (MAA Notes no. 26, pp. 14–25). Washington, DC: Mathematical Association of America.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

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

Pollatsek, A., Lima, S., & Well A. D. (1987). Concept of computation: Students' understanding of the mean. *Educational Studies in Mathematics, 12*, 191–204.

Rosebery, A. S., & Warren, B. (1992). Appropriating scientific discourse: Findings from language minority classrooms. *Journal of the Learning Sciences, 2*(1), 61–94.

Roth, W.-M., & Bowen, G. M. (1994). Mathematization of experience in a grade 8 open-inquiry environment: An introduction to the representational practices of science. *Journal of Research in Science Teaching, 31*, 293–318.

Roth, W.-M., & McGinn, M. K. (1997). Graphing: Cognitive ability or practice? *Science Education, 81*(1), 91–106.

Russell, S. J., Schifter, D., & Bastable, V. (2002a). *Developing mathematical ideas: Working with data.* Parsippany, NJ: Dale Seymour Publications.

Russell, S. J., Schifter, D., & Bastable, V. (2002b). [Fourth-graders representing data of height on stacked dot plot.] Unpublished raw data.

Russell, S. J., Schifter, D., & Bastable, V. (2002c). [Fifth-grade students comparing graphs.] Unpublished raw data.

Scheaffer, R. L., Watkins, A. E., & Landwehr, J. M. (1998). What every high-school graduate should know about statistics. In S. P. Lajoie (Ed.), *Reflections on statistics: Learning, teaching, and assessment in grades K–12* (pp. 3–31). Mahwah, NJ: Erlbaum.

Schwartz, D. L., Goldman, S. R., Vye, N. J., & Barron, B. J. (1998). Aligning everyday and mathematical reasoning: The case of sampling assumptions. In S. P. Lajoie (Ed.), *Reflections on statistics: Learning, teaching, and assessment in grades K–12* (pp. 233–273). Mahwah, NJ: Erlbaum.

Sconiers, S. (Ed.) (1999). *Bridges to classroom mathematics: Mathematics handbook.* Lexington, MA: COMAP.

Shaughnessy, J. M. (1992). Research in probability and statistics: Reflections and directions. In D. Grouws (Ed.), *Handbook of research on the teaching and learning of mathematics* (pp. 465–494). New York: Macmillan.

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

Tufte, E. R. (1983). *The visual display of quantitative information.* Cheshire, Connecticut: Graphics Press.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

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

Watson, J. M., & Moritz, J.B. (2000). Developing concepts of sampling. *Journal for Research in Mathematics Education, 31*, 44–70.

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