STRATEGIES FOR MANAGING STATISTICAL COMPLEXITY WITH NEW SOFTWARE TOOLS

JAMES K. HAMMERMAN AND ANDEE RUBIN
TERC
Jim_Hammerman@terc.edu; Andee_Rubin@terc.edu

SUMMARY

New software tools for data analysis provide rich opportunities for representing and understanding data. However, little research has been done on how learners use these tools to think about data, nor how that affects teaching. This paper describes several ways that learners use new software tools to deal with variability in analyzing data, specifically in the context of comparing groups. The two methods we discuss are 1) reducing the apparent variability in a data set by grouping the values using numerical bins or cut points and 2) using proportions to interpret the relationship between bin size and group size. This work is based on our observations of middle- and high-school teachers in a professional development seminar, as well as of students in these teachers' classrooms, and in a 13-week sixth grade teaching experiment. We conclude with remarks on the implications of these uses of new software tools for research and teaching.

***Keywords*:** Representations; Software tools; Variability; Proportional reasoning; Group comparison; Covariation; "Binning"

1. OVERVIEW

This paper reports on research at the intersection of two lines of inquiry: 1) What approaches do people use to deal with variability in data? and 2) How do new statistical visualization tools change the strategies people use in analyzing data? Each of these questions is worthy of study in its own right. Variability, while at the heart of statistics, presents a significant challenge to teachers and students trying to develop a sophisticated set of statistical reasoning strategies. Its ubiquitous presence in data makes simple statements that take *all* the data into consideration impossible, unless one can somehow acknowledge and "tame" the variability by working with fewer numbers. In some ways, a single measure such as the mean is the ultimate way to deal with variability in a distribution, since its role is, in fact, to reflect all the values in a data set with just one number. However, values such as the mean are notoriously difficult for students and teachers to understand as representing the entire data set at once (Konold, Higgins, Russell, & Khalil, 2003; Konold, Pollatsek, Well, & Gagnon, 1997; Konold et al., 2002; Mokros & Russell, 1995). So there is a need for other ways to deal with variability that teachers and students can understand and appropriate.

Methods for handling the variability in data depend intimately on the tools at hand, and therefore new software visualization tools are dramatically changing the way data analysis and statistics are learned and taught. Until recently, and to some extent still, the techniques most students and their teachers learned for describing a data set were primarily numerical, i.e., computing measures of center such as the mean or median, and computing measures of variability such as the standard deviation or inter-quartile range (IQR). The five-number summaries illustrated by box plots went one step further in describing a distribution by including both central tendency and some indications of variability. All these numerical characterizations have been made easier to obtain by the accessibility of calculators, often with built-in statistical functions, so that any student is able to compute some basic statistics about a distribution, even if she/he doesn't really know what they mean. With these current tools, computations in general are easy, as are some rudimentary graphical manipulations, but the new kind

Statistics Education Research Journal 3(2), 17-41, http://www.stat.auckland.ac.nz/serj

of tool we discuss in this paper greatly expands the possible ways teachers and students can interact with data.

Using any new tool or representation necessitates change in the content and pedagogy of statistics instruction and in many cases teachers are unprepared for these changes. Even the primarily pencil and paper data displays developed to simplify the task of visualizing a distribution in the past 30 years (Tukey, 1977) can present a challenge to teachers. Dot plots, stem and leaf plots, and box plots are particularly popular in elementary and middle school textbooks, and classroom conversations about data are now expected to include more nuanced descriptions of distributions that include the shape of the data, including issues of center, variability, skewness, and other characteristics best uncovered with graphical representations. Many teachers, however, have limited experience with these new approaches, so the potential for deeper classroom discussion afforded by these representations may be lost. New interactive visualization tools, such as the one described in this paper, have the potential to support further changes in classroom approaches to data analysis, and thus present yet another challenge to teachers. To effectively teach students about data analysis, teachers will have to both know the mathematical content and understand students' statistical reasoning as it develops. Because of these new classroom responsibilities, we have worked primarily with teachers, as mediators of student learning.

1.1. THE POTENTIAL EFFECTS OF NEW INTERACTIVE VISUALIZATION TOOLS

New interactive visualization tools (e.g., Tabletop™ (Hancock, 1995), Fathom™ Dynamic Statistics™ (Key Curriculum Press, 2000), and TinkerPlots™ (Konold & Miller, 2004)) create yet another set of new possibilities both for data display and for attaching numbers to graphs in informative ways. No longer are we limited only to numbers that we can calculate from formulas, often without looking at the distribution itself, and to a few static graphs. People using these software tools easily see the shape of data and other distributional characteristics that numerical summaries alone can't show without constructing graphs by hand. More important, they can dynamically manipulate data displays, and begin to get a feel for the data themselves and for relationships among data attributes. These software tools make possible a pedagogical change that embodies a view of data analysis that allows and even encourages students to transform graphs interactively and to request a variety of numbers that represent the distribution as it is displayed in the graphs.

These new tools by themselves don't reduce the complexity of data analysis, nor do they solve the problem of variability. People still have to attend to what can be an overwhelming amount of information, i.e., all the data values, although visualization tools hope to take advantage of our built-in perceptual apparatus to ease this task. In addition, these tools give people more options for describing distributions in ways that can be useful in statistical decision-making. The affordances these new tools provide and how people use them have been explored to some extent, primarily by developers of the particular tools and their collaborators (Bakker & Gravemeijer, 2004; Cobb, 1999; Hancock, Kaput, & Goldsmith, 1992; Konold, 2002; Konold et al., 1997; Rubin, 1991, 2002; Rubin & Bruce, 1991). However, these are only preliminary studies and barely scratch the surface of the inherent flexibility and complexity of using these tools. Researchers and teachers need more detailed information about what happens when learners analyze data with this kind of software over an extended period of time.

Interactive software data visualization tools which allow for the creation of novel representations of data open up new possibilities for students (and teachers) to make sense of data, but also place new demands on teachers to assess the validity of the arguments that students are making with these representations, and to facilitate conversations in productive ways (cf. Ball, 1993; Ball, 2001). Just as teachers in general need a deeper understanding of the mathematics they teach to work effectively in reform-oriented classrooms (Ball, 1991; Ball, Hill, Rowan, & Schilling, 2002; Borko & Putnam, 1995; Hill & Ball, 2003; Russell et al., 1995; Schifter, 1997; Sowder, Philipp, Armstrong, & Schappelle, 1998), so, too, will teachers need a deeper understanding of data and statistics (Rubin & Rosebery, 1988; Russell et al., 2002) to use new software tools in their classrooms. Professional development for teachers will need to address issues of mathematical content, as well as issues of learning, representation, and pedagogy. By exploring and discussing data for themselves in new ways,

teachers can develop a deeper understanding of the mathematics, and also of how classroom discourse and pedagogy might change through use of new software tools. However, *teachers'* experiences of learning with these new software tools have not yet been explored.

Many of the teachers came to our professional development seminar thinking that statistics is about mean, median, and mode. They knew how to calculate these statistics, though they didn't always have robust images about what they meant or how they were used. In general, they had not dealt with data sets that required them to confront significant variability, so that they didn't have strategies to apply to the new complexity they encountered using interactive visualization tools to explore real data sets. Finding that there was a lot more to see and analyze in data than they had ever imagined was both powerful and intimidating. We were interested in how these teachers, with their diverse prior skills and experiences in data analysis, represented and made arguments about data. Although our work was primarily with teachers, the strategies and approaches we saw them use were similar to those students used in the classrooms we observed.

This paper, then, describes research that primarily involved teachers who were learning to analyze data in a professional development seminar as well as students in several middle school classes. Our research sought to answer two questions: 1) What statistical reasoning strategies do teachers employ to handle issues of variability when analysing data? and 2) What new affordances does a tool such as TinkerPlots™ provide for coping with variability? We will describe ways that students and teachers used the new options available to them in TinkerPlots™ (Konold & Miller, 2004) to compare groups. In this context, there is almost never a way to make a clear judgment by simply producing a picture or two, except in the rare instance when there is no overlap between the distributions to be compared. How does a choice of representation and measure help simplify the difficult task of making a decision about two distributions which each have significant variability? We will describe two main approaches that are made possible by TinkerPlots™—those that use categorizing or "binning" continuous data to reduce the apparent variability; and those that use proportional reasoning, primarily to deal with issues of unequal group sizes. We describe several examples of ways in which teachers and students with access to TinkerPlots™ use each approach, comment on the validity of each technique, demonstrate how each approach attempts to confront and handle some of the complexity and variability inherent in data, and comment on the developmental course of the use of some of these strategies.

1.2. A PERSPECTIVE ON LEARNING

Our view of learning influenced a variety of aspects of our research: our choice of format for our teacher seminars, our choice of data sets and discussion topics, and our conclusions about teachers' and students' reasoning. Because learning to analyze data is a process that unfolds over time, we believe that we need a series of rich tasks and deep discussions to understand teachers' approaches and how they develop. Although we offered a data analysis seminar for teachers, our focus was less on teaching them specific concepts or techniques than it was on providing an environment in which teachers could explore important ideas about data and statistics using new software tools, and on conducting research on their thinking. Our teaching was, thus, strongly constructivist and our teaching and research were closely integrated as we asked questions and facilitated discussions focusing on the different ways that teachers were making sense of problems (cf. Duckworth's (1996) notion of "Teaching as Research"). Teachers often shared the several ways they made sense of particular problems, and discussions served to clarify these different approaches through questioning and explanations. Not all disagreements were resolved immediately, but teachers seemed comfortable letting confusing ideas simmer and re-emerge over the course of several weeks.

From this perspective, it is also important that the data sets and problems we provided were complicated enough to make important issues in data analysis salient. For example, if the data sets that learners are given to compare always have the same number of cases in each group, they will never have to confront the issue of unequal group size, or to think about ways to mathematically "equalize" the numbers in each group—they'll get the same results whether they compare groups using absolute numbers or percentages. It is only when the groups are *different* sizes that the

difference between using numbers or percentages becomes relevant. More complicated data sets, messier ones, are both more realistic and can give learners a chance to grapple with some important statistical ideas, though increased complexity may sometimes overwhelm novice learners. Helping learners manage the confusion that comes with this complexity so that it doesn't overwhelm them is an important part of our teaching.

Finally, we believe that learning is a slow, non-linear process of constructing and building more and more robust understandings over time. Powerful ideas, especially those requiring a move towards abstraction, may need to be built and re-built several times before they become solid and can be used in a wide variety of situations. They may appear to be in place at one point in time, but later, can appear to be "lost" when they are needed in a different, more complex context. The example of sixth graders embracing and then questioning the use of percentages (described in section 4.2) illustrates this. This view of learning is an important lens to use in reading the descriptions that follow, as teachers and students may appear to be reasoning inconsistently. Iterative building and rebuilding of ideas, we claim, is one of the hallmarks of the learning of important and difficult concepts.

2. REVIEW OF THE LITERATURE

Some of what makes working with data complex is the tension between attending simultaneously to individual values and to aggregate properties of distributions. While expert data analysts move fluidly between appropriate attention to these different levels of understanding data, this is a difficult perspective to attain. Several other views of data typically precede seeing it as a distribution with aggregate properties. One common perspective is to view data as a collection of individual values without any relationship to one another (Bakker & Gravemeijer, 2004; Hancock et al., 1992).

Konold and colleagues (Konold & Higgins, 2002; Konold et al., 2003) argue that children see data in *several* simpler ways before ever noticing aggregate and emergent features of data sets. Their fourfold schema includes the following different ways of viewing data, which we consider useful for examining the thinking of adults as well as children:

1. Data as a *pointer* to the data collection event but without a focus on actual data values—in this view, data remind children of their experiences, "We looked at plants. It was fun."
2. Data as a focus on the identity of individual *cases*—these can be personally identifiable, "That's my plant! It's 18 cm tall," extreme values, "The tallest plant was 37 cm," or interesting in some other way.
3. Data as a *classifier* which focuses on frequencies of particular attribute values, or "slices," without an overall view—"There were more plants that were 15 to 20 cm than 10 to 15 cm."
4. Data as an *aggregate*, focusing on overall and emergent characteristics of the data set as a whole, for example, seeing it as describing variability around a center, or "noise" around an underlying "signal" (Konold & Pollatsek, 2002)—"These plants typically grow to between 15 and 20 cm."

It is possible to make group comparisons from any of the case, classifier, or aggregate perspectives, i.e., comparing extreme values, comparing the numbers in particular slices, or the more canonical comparison of means, respectively. However, aggregate views are preferable, as they are required to look beyond the data towards making inferences about the underlying populations or processes represented by data samples. Konold and Pollatsek (2002) argue that it is sometimes easier to use aggregate measures of center when comparing groups than when looking at data distributions on their own. When comparing groups, they claim, it is clear that the focus is on underlying processes rather than the particulars of the data at hand, although earlier work (Konold et al., 1997) found that students had difficulties thinking of underlying "propensities" even when comparing groups. In the work reported here, both teachers and students only rarely used formal measures of center to characterize data sets even when comparing groups. Still, among the several methods we will report, there were some that demonstrated aggregate thinking without involving the use of measures of center.

As we and others suggest, data are most interesting when they are used to make inferences beyond themselves, that is, when they are seen as representative of a larger population about which one wants to generalize. This process of generalizing from a sample to a population is notoriously difficult. When comparing groups with data seen as a sample, the inherent variability of a particular set of data is complicated further by the fact that we must also determine whether observed differences in data reflect underlying differences in populations, or are merely due to chance fluctuations in a sample. The difficulty that people have in understanding the relationship between sampling variability and the inherent variability of the underlying population is well documented (Rubin, Bruce, & Tenney, 1990; Saldanha & Thompson, 2002; Sedlmeier & Gigerenzer, 1997; Watson & Moritz, 2000). Yet, while issues of sampling variability sometimes arose for teachers and students in this study, such variability is not the focus of this paper. We will, however, point to instances when issues of sampling were salient.

The TinkerPlots™ software we used in this study made it easy to divide data into "bins" in which cases within a range of values of an attribute are grouped together. The impact of such a representation has been little explored. Cobb (1999) reports how students using his minitools (Cobb, Gravemeijer, Doorman, & Bowers, 1999) were able to partition data in equal sized groups, allowing them to make arguments about the position of the middle half of the data (using a representation akin to box plots); and were able to partition data into groups with a specified interval width, allowing for arguments about the numbers or percentages on either side of a fixed value. Cobb, McLain and Gravemeijer's (2003) 8th grade (age 13) study argues for the utility of breaking bivariate data into a series of distributions of equal width "slices" of the independent variable in order to look for patterns in the position of these distributions and their means across the slices, that is, splitting bivariate data into a series of group comparisons which, they argue, is conceptually (if not actually) what expert data analysts do when looking for a regression line. Meletiou and Lee (2002) describe difficulties that students have with histograms, another form of grouping data, stating that "the research literature tells us very little about how understanding of histograms and other graphical representations develops" and calling for further research. Finally, some studies argue that students often like to characterize data by "hills" (Cobb, 1999), "modal clumps" (Konold et al., 2002) or "bumps" (Bakker, 2004), that is, central slices of a distribution containing a large proportion of the data, though these categorization schemes rely on natural breaks in the shape of a data distribution rather than equal width "bins".

When people use the "binning" features of software, they typically describe what they're seeing either by general comments on the shape of data, by comparing the number of data points in different bins, or by comparing the percentage of data points in different bins. The multiplicative reasoning, including proportional reasoning, needed to use the percentage strategy is important to thinking well about data, and it has been highlighted by several researchers. Shaughnessy (1992) in his review article claims that the ratio concept is crucial in statistics and often lacking among students (p. 479). Upon re-examining their comprehensive framework for middle school students' statistical thinking (Mooney, 2002), Mooney and colleagues (Mooney, Hofbauer, Langrall, & Johnson, 2001) changed just two elements—they modified the category for organizing and reducing data, and added the important missing element, multiplicative reasoning, which they describe as, "reasoning about parts of the data set as proportions of the whole to describe the distribution of data or to compare data sets" (p. 438). Saldanha & Thompson (2002) argue that seeing a sample as a "quasi-proportional small-scale version of a larger population" is an important conceptual move for students in making statistical inferences. In a 7th grade (age 12) teaching experiment, Cobb (1999) proposed, "our goal for the learning of the classroom community was that reasoning about the distribution of data in multiplicative terms would become an established mathematical practice that was beyond justification" (p. 11) and described a fair amount of reasoning by use of "qualitative proportions" in their analysis. In a subsequent 8th grade (age 13) teaching experiment focusing on looking for patterns in bivariate data, Cobb and colleagues (Cobb et al., 2003) took multiplicative reasoning as the starting point. In fact, the multiplicative reasoning needed to normalize data, that is, to make the scale of numbers the same so they can be compared, is a powerful technique used widely throughout statistics. Examples include rescaling variability in standard deviation units when calculating Z-scores, calculating the mean to yield a per case measure of an attribute, or using percentages instead of counts to deal with differences in sample size, as we will see in this study.

While important in statistics (and elsewhere), proportional reasoning can be difficult, especially for students who are attempting to distinguish it from additive reasoning (Harel & Confrey, 1994). Lamon (1994) details some of this complexity that is especially relevant for data analysis, stating that proportional reasoning requires "unitizing", i.e., "the ability to construct a reference unit or unit whole, and then to reinterpret a situation in terms of that unit" (p. 93), as well as "norming" which includes the idea of percentages, i.e., "reconceptualizing a system in relation to some fixed unit or standard" (p. 94). These transformations require shifting attention from units to relationships among units, a more abstract idea. At the same time, working with these relationships reduces the amount of data to which one must attend at the same time which, Lamon argues (citing Case (1978; 1980)) may "facilitate reasoning [by] easing the load on the working memory" (p. 112).

This paper describes our experiences studying teachers' and students' uses of binning and proportional reasoning strategies using a computer tool that makes each of these strategies more accessible and flexible. Since such tools are new and not yet widely disseminated, we know very little about how teachers' and students' strategies develop when they have these resources available. Which components of the software do teachers use and how do they take advantage of the interactive possibilities afforded by the software? What can we learn about teachers' and students' statistical reasoning by analyzing their interaction with these new tools? What implications can we draw from these data for teaching and learning?

3. CONTEXTS AND METHODS

The data for this paper come from several sources, all connected with the Visualizing Statistical Relationships (VISOR) project at TERC in Cambridge, Massachusetts, USA. VISOR is a teacher professional development and research project studying how people learn about data analysis and statistics and how computer visualization tools can enhance that learning. In VISOR, the professional development and research goals were often mixed. We offered opportunities for teachers to explore data topics such as ways of describing data, stability of measures and the role of sample size, making inferences about group comparison and co-variation situations, and confidence intervals, among others. However, we focused less on *teaching* teachers specific things than on exploring their thinking in the context of use of computer software tools. Teachers explored a variety of data sets using two innovative software tools, TinkerPlots™ (Konold & Miller, 2004) and Fathom™ Dynamic Statistics™ (Key Curriculum Press, 2000). In the group, they also talked about teaching about data analysis, and brought in examples from their own classrooms of their students' thinking and work using these tools. By focusing on how people think about and explore data, the project hoped to help teachers develop a sense of themselves as data analysts, to understand better some of the issues that arise in learning about data and statistics, and to feel more confident teaching about data in richer and deeper ways.

In the VISOR seminar, we worked with a group of 11 middle- and high-school teachers (8 women, 3 men; 6 middle school, 5 high school; 10 White, 1 Black) from mostly urban Boston-area schools, meeting biweekly for three hours after school over the course of two years. In fact, only eight of the teachers continued into the second year. In its third and final year, VISOR worked with a new group of nine teachers. Teachers varied in their comfort with computers and in their prior experience with statistics, some had had very little exposure, a few taught AP Statistics (Advanced Placement high-school courses that provide college credit) or had done data analysis in industry. While some taught semester- or year-long statistics courses, most only taught about data and statistics during a few weeks each year.

We videotaped group sessions, took extensive field notes, and collected teachers' work from the seminar, including video feeds from the computer work of one small group each session. After each session, we created a rough log of the videotape, developing more detailed transcripts for certain sections as needed in our analytic work. We also observed teachers in their classrooms, and collected field notes and examples of students' work, as well as copies of what teachers brought in from their classrooms. Several times during the two years, we conducted formal, individual, audio- or videotaped interviews with teachers on a variety of topics. Finally, one of us conducted a 13-week

teaching experiment on data analysis with a group of 12 relatively advanced sixth grade students (age 11–12) from an urbanized suburb of Boston, taking field notes and reflective teaching notes after each session, and also collecting examples of student work.

Our research goals and methods were primarily descriptive and exploratory, within the goal of discovering how teachers used new capabilities of TinkerPlots™ to compare groups. The authors, who collaboratively led the seminar as well as the research, met regularly to discuss teachers' prior work, to puzzle through what the data showed about how different teachers were making sense of the problems, and to plan sessions to illuminate and highlight different conceptions. These analyses most resembled a combination of group clinical interview (Clement, 2000) and teaching experiment methodologies (Steffe & Thompson, 2000). The authors were sometimes joined in these discussions by Bill Finzer (designer of Fathom™) and Cliff Konold (designer of TinkerPlots™) to focus on aspects of the software tools that might affect teachers' thinking. We also met regularly with a research team at TERC to more closely analyze the formal, transcribed interview data. In this process, pairs of researchers separately analyzed each transcript using a combination of etic codes developed from our theoretical frameworks, and emic codes that emerged from the interviews themselves (Miles & Huberman, 1984; Patton, 1990). We wrote memos about each participant and discussed discrepancies in our analyses until we reached agreement. We then compared across participants to look for common themes and methods, as well as for interesting variations.

3.1. SOFTWARE TOOLS

While we used both TinkerPlots™ and Fathom™ in VISOR sessions, the TinkerPlots™ software provided the platform for the examples we will use in this paper. TinkerPlots™ is a data analysis environment primarily for middle-school classes that provides students with a wide range of tools to create traditional and non-traditional data representations. By various combinations of sorting and separating data into categories, ordering or highlighting information by the value of an attribute, and stacking and otherwise organizing data, users can make graphs that are both familiar and very different from those typically seen in school or, for that matter, in most research settings. Users can display numerical information about plots: the value and position of the mean, median, mode, or midrange; the number or percentage of cases in bins or between sets of moveable dividers; or the value at a moveable horizontal or vertical line. The software offers tools for displaying data as value bars (in which the length of each bar represents the magnitude of the value of an attribute for a case), or fused into rectangular or circular areas. Finally, it provides tools for creating box plots, as well as innovative "hat plots" that partition the data like box plots, but based on user specifications such as percentages of the range or of the data, or numbers of standard or average deviation units from a center.

3.2. DATA SETS

While our work with teachers and students involved exploring a wide variety of data sets, the results we present in this paper focus (primarily) around two data sets. The first, explored in the middle of the first year, was invented but realistic data, modified from Cobb et al. (1999), comparing the efficacy of two drug protocols for treating patients with HIV-AIDS. Of the 232 patients in the sample (160 men and 72 women), 46 randomly received an Experimental protocol and 186 received the Standard protocol. Outcomes were measured in patients' T-cell blood counts and teachers were given information from an AIDS education group stating that normal counts ranged from 500 to 1600 cells per milliliter. We added a gender attribute to the original data designed in such a way as to show an interaction between gender and protocol in their effects on T-cell counts, that is, differences across categories and an interaction of categories in a numerical attribute.

The second data set, explored early in the second year, consisted of real survey data of 82 students (34 girls and 48 boys) attending two western Massachusetts high schools (51 from Holyoke and 31 from Amherst) collected by Cliff Konold in 1990 and included with TinkerPlots™ (US Students: Konold & Miller, 2004). Attributes include students' height and weight, number of older and younger

siblings, hours spent doing homework and working, and average grades received, among others. Teachers focused on school and gender differences in grades received and on hours per week spent doing homework, that is, differences across categories in a categorical and a numerical attribute.

We will also look briefly in this paper at two data sets exploring the relationship between two numerical attributes, both of which were explored late in the second year of VISOR. The first compares the median age in each state with the percent of the state population voting for George W. Bush in the 2000 U.S. Presidential election. The second looks at the relationship between average state educational spending and number of teen births per 100,000 population in the U.S.

These data sets offering different types of data and relationships among data allowed us to see a range of ways that teachers and students thought about and dealt with the variability in data using TinkerPlots™.

4. RESULTS AND DISCUSSION

We discuss in this section results pertaining to our research questions and the relationships between them. We describe how teachers with diverse statistical analysis and teaching experiences approach issues of variability, especially when comparing groups with one numerical and one categorical variable. We also document in less detail examples of looking at the relationships between two numerical variables. Throughout this section, we also relate these findings to our second research question, regarding the affordances offered by a statistical visualization tool like TinkerPlots™.

The purpose of describing *teachers'* use of TinkerPlots™ in comparing groups is two-fold: First, it suggests the kinds of strategies that *students* might use as well when they have TinkerPlots™ as a resource. In fact, as described below, we have confirming evidence from observations in classrooms that students and teachers share these approaches. Second, it helps us learn how teachers approach comparing groups tasks and, therefore, what they need to learn to guide a statistically meaningful conversation for their students.

The teachers in the VISOR seminar created many previously unseen (at least by us) graphs and were extremely creative in their approaches to comparing groups in the data sets described above. In general, our results confirmed Konold et al.'s (1997) observation that students (in this case teachers) seldom use a measure of center as their first method for comparing two data sets presented in graphical form. We report here on two key types of strategies that teachers used in comparing groups with TinkerPlots™, describe how essential design features of TinkerPlots™ influenced these strategies, and comment on whether and how these techniques developed as the VISOR seminar went on. In addition, we describe evidence that students' use of TinkerPlots™ brings up many of the same data analysis issues that arose in the teacher seminar. We situate these descriptions in the general dilemma teachers and students are facing: How to make comparisons between two groups when each of them exhibits considerable variability.

The primary strategy we analyze is one that is both new and unusually easy to use in TinkerPlots™, analyzing a distribution by dividing and chunking it into several pieces. While many statistical analysis packages provide ways to create histograms and even to change their bin sizes in a limited way, it is the flexibility that TinkerPlots™ provides for manipulating and observing information about sections of a distribution that creates a new and powerful tool for learners. Dividing a distribution into bins is effortless in TinkerPlots™, as this representation is a primitive among the graph operations TinkerPlots™ makes available. The strategy of creating multiple "bins" along an axis effectively reduces the number of actual values in the distribution, thus reducing apparent variability.

The need for the second strategy we describe arises in some part from using the binning strategy in the context of comparing groups. If any two groups are of unequal sizes, the binning process will require an understanding of proportional reasoning to compare the two groups. In this context, we see both teachers and students struggling with the difference between additive and multiplicative reasoning. The first pays attention to the *number* of points in a bin, while the second focuses on the *proportion* of points in each bin. TinkerPlots™ supports both of these kinds of reasoning in slightly different ways. We analyze here the ways in which TinkerPlots™' features affect students' and

teachers' uses of additive and multiplicative reasoning and explore the new strategies and dilemmas that arise from the use of TinkerPlots™ features.

4.1. ANALYZING DATA IN BINS

One of the most common techniques both teachers and students used to compare groups consisted of limiting the number of unique values in the data sets by creating two or more bins. Considering a set of data points as a relatively small number of groups of values is not a new idea; in fact, it is the basis for many statistical representations. Histograms and box plots both partition data sets into groups within which all values are essentially the same for analytical purposes.

The role of bins in TinkerPlots™ is notable because the software automatically goes through a "binned" stage as data are separated according to the value of a highlighted attribute; the user does not need to imagine or specifically request a graph with bins. Thus, many more of the graphs produced in our seminar used bins than might have been the case with other software tools. This representational strategy is not specific to TinkerPlots™. In fact, we have seen similar binning approaches when teachers and students analyze data with pencil and paper. But the immediacy of TinkerPlots™' binning functions affords more complex strategies. We describe below some of the common ways teachers and students used bins to compare groups.

Each of the methods described below highlights some of the consequences of regarding data in bins. They are all examples of a tension inherent in data analysis; finding the right balance between the advantages of reducing variability as a way to deal with the complexity of data, on the one hand, and the risks of making claims that might not be true, or would have to be qualified, were all the data included in the analysis, on the other. Several of the examples below also illustrate the interplay of contextual and strictly numerical ways of looking at data, which we frequently observed in teachers' discussions of their analyses.

Using cut points, both system- and user-generated

One of the simplest ways to create bins in a distribution is to divide it into two parts. We have used the term "cut point" to designate a value in a distribution which divides it into two groups above and below that point. When a user begins to separate the values of a variable, TinkerPlots™ immediately provides a single cut point using the software's rule of using a value roughly at the rounded midrange.

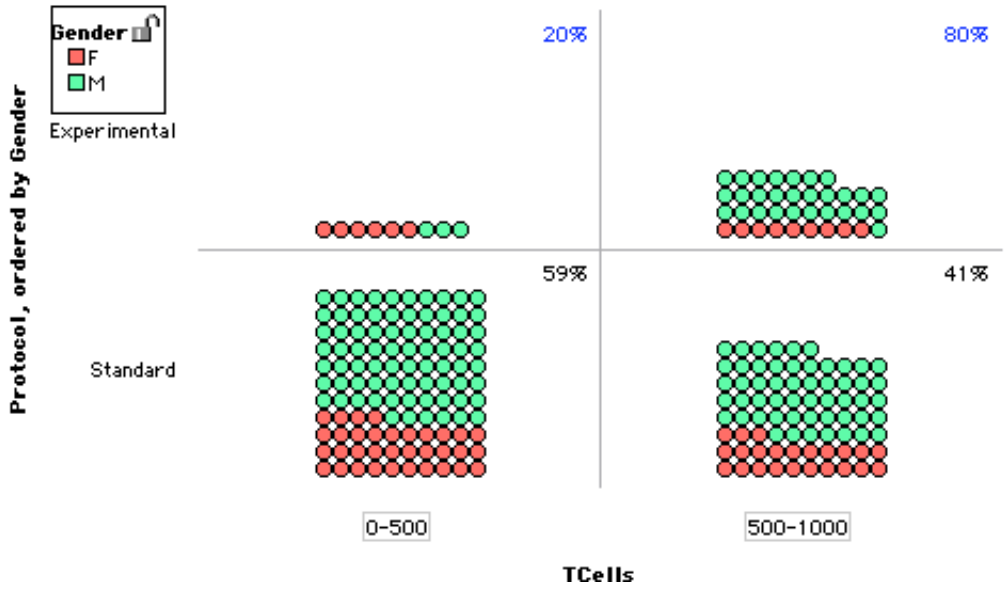

Figure 1. Comparing the percent of each protocol above a single cut point

If the data are split into two subgroups as well, we get a two-by-two representation (see Figure 1). A legitimate comparison between the two groups, then, would involve comparing the percentage of each group above and below the bin boundary. Understanding the importance of using percents rather than counts is an issue we discuss in Section 4.2, below. In fact, TinkerPlots™ supports this comparison by making it possible to display the percents in each bin. In Figure 1, an argument can be made for the superiority of the experimental treatment based on the larger percentage of experimental cases in the high T-cell bin. Remember, high T-cell counts are better.

While this kind of representation was common early in our seminar because it was so easy to create, teachers quickly grew to reject this representation because it hid so many of the details of the shape of the data. A more common representation was Figure 2. Here we see that the teacher created seven bins and displayed the percents in each. In this particular case, the analysis proceeded by adding together the percents in the bins above 500 T-cells per ml for each protocol, replicating the analysis in Figure 1. At other times, however, these kinds of multi-bin graphs were analyzed in very different ways, as described below under "Comparing slices."

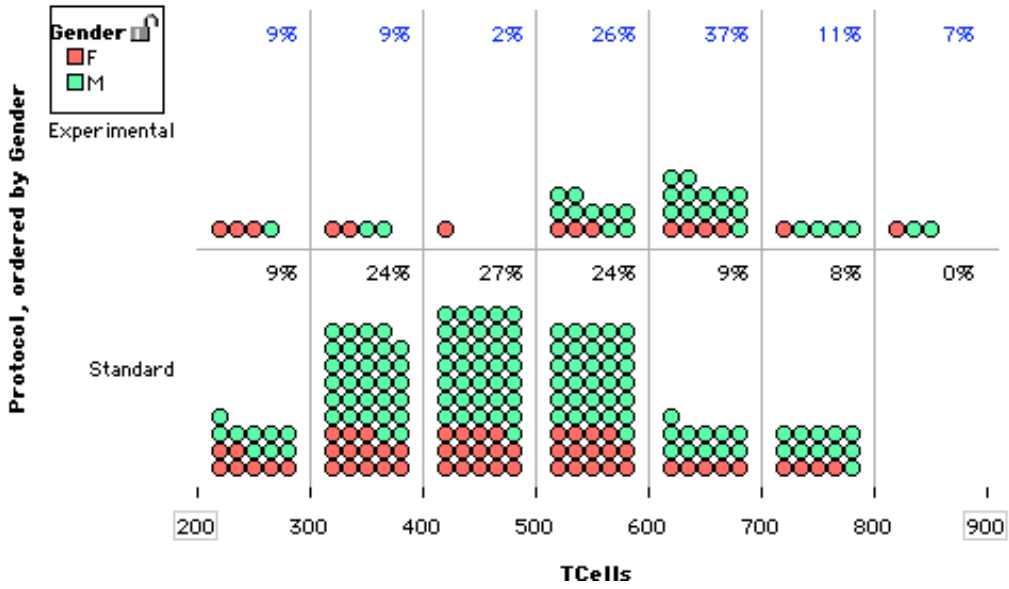

Figure 2. AIDS T-cell data in seven bins with row percents

A more sophisticated approach using TinkerPlots™ was to use a different tool, *dividers*, that allows the user to specify exactly where she wants to put one or several cut points. In Figure 3, the teacher has placed the cut point at exactly 500 using a divider tool to split the distributions into two parts. This representation is different from a binned representation in that the user has to be explicit about creating and moving a divider (by grabbing the small white square at the right side of the page). Note that in this case, the default setting at which TinkerPlots™ drew the first bin and the cut point value that this teacher has chosen are the same: the difference is that the teacher has detailed control of the dividers and could change the cut point to 495 or even 497.

Representations with dividers arose fairly early in the seminar, although it remained difficult for some teachers who had trouble mastering the tool and therefore continued to use primarily bins in their analyses. This representation is interesting because, while imposing a cut point, it also retains visual information about the rest of the shape of the distributions. Unlike Figures 1 and 2, the exact T-cell value of each point is represented on the graph. This became a preferred graph for some teachers who would routinely "separate a variable completely" (create a representation without bins), then impose their own dividers. Figure 3 is also different from the rest of the graphs in this paper because the continuous variable is displayed vertically, but that is not unusual in TinkerPlots™, since it is equally easy to put a variable on either axis.

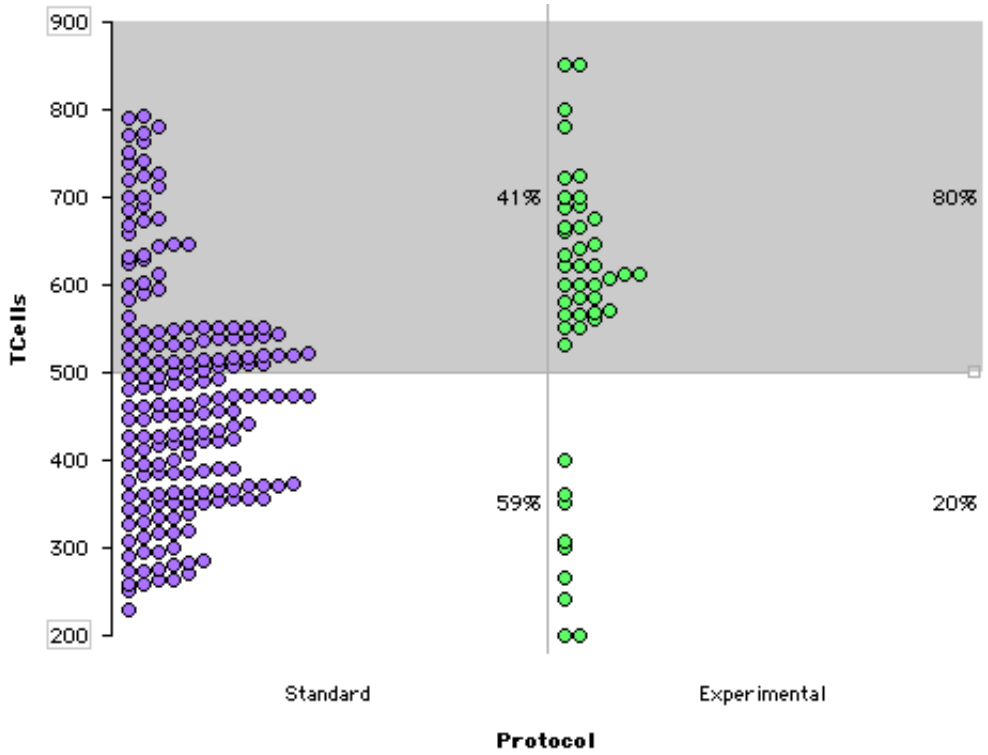

Figure 3. Using a divider to compare percentage of T-cells above and below a user-specified cut point

This description of three graph types is not intended as a developmental sequence, as different teachers made different choices in representations at any point, and an individual teacher might have alternated among several kinds of representations. There are, however, a few general points to be made about trends in teachers' choices.

Most teachers did prefer Figure 2 over Figure 1 (at least, after the first week or two), even though they performed essentially the same analysis with each: adding up the percentage of cases with values over 500. In fact, using Figure 2 to make the argument requires more work than Figure 1, since the percents above 500 are not yet totaled. Our observations lead to this hypothesis: The teachers used Figure 2 so that they could CHOOSE the cut point based on the shape of the data, then use their chosen cut point to compare percents across groups with regards to that cut point. Working with a cut point from Figure 1 runs the risk of choosing a value that would seem more or less representative of the data depending on the distributional shape. Figure 2 combines the ease of using bins with a desire to see the shape of the data. In this case, we hypothesize that if the distribution were less symmetrical, teachers may have wanted to use a different approach than the single cut point one. Some tentative findings from research not reported here support this hypothesis.

What influenced a teacher to choose between Figures 2 and 3 in analyzing a data set? As described above, teachers did not move from using only a graph like Figure 2 to using only a graph like Figure 3. However, it often happened that when teachers "completely separated" a graph, such as that in Figure 3, they then re-binned the data (which one can do easily in TinkerPlots™). Our hypothesis about this tendency is that, faced with a large number of points stacked unsteadily on the axis, teachers often chose to retreat to a representation with more built-in structure and less apparent variability. Interestingly, in none of these examples were teachers likely to look at a measure of center such as mean or median.

All of these graphs were generated as part of the teachers' discussion of the AIDS data, which took place after three months of the seminar. In generating and making arguments from graphs such as Figures 1 through 3, teachers often used the value 500 as a cut point, since that was the lower end of the normal range of T-cell counts according to the information sheet we had given them. Using

these representations, teachers argued in support of the Experimental treatment's superiority by noting that the percentage of patients above 500 was greater in the Experimental group than in the Control group. Using graphs such as Figure 3, teachers noted that, "The Experimental treatment yields a higher percentage of participants in the normal range (above 500). [Figure 3] shows that in the Experimental protocol, 80% of participants were in the normal range, while in the Standard protocol, only 41% of participants had T-cell counts above 500." One teacher was more effusive, saying, "A huge preponderance don't get much better than 500 in the Standard, though a huge preponderance do better than 500 in the Experimental."

Exploring these three classes of representations led to a discussion in the group about the rigidity and arbitrariness of the cut point, a discussion that included both contextual and statistical arguments. Some teachers wondered how rigid the value of 500 was from a medical perspective, would T-cell counts of 495 mean someone was *nearly* healthy? What about 450? One teacher made an analogy with grouping people by height saying, "If somebody was right at the borderline that doesn't mean they're short." Other teachers speculated that there *could* be a biological mechanism that created a sharp distinction in sickness rates once T-cell counts reached a specified level. "There may be a magic number that's very close. To get infected by a germ, there has to be a minimum number. If you get less than the minimum number, you get nothing." Teachers also made arguments for the significance of the cut point 500 grounded in the shape of the distribution. Some people noticed that there was a large gap in the data for the Experimental group just below 500, which gave them confidence in using this value to divide the data. This, in turn, led to further context-driven speculation about a possible genetic difference in response for the people at "the low end of the Experimental protocol." The back and forth between patterns observed in the data and those driven by the context led the participants beyond what they would have seen by just looking at the distribution as numbers. After they had all seen the entire distribution, most teachers seemed comfortable with reducing variability by imposing a single cut point, although there was some disagreement about exactly where to put it, thus focusing on this one distinction rather than trying to integrate and deal with the variability of the entire data set at once.

In some ways, all of these uses of cutpoints can be seen as examples of viewing data as a classifier using Konold et al.'s (2003) schema, i.e., attending only to the frequency of data in a particular bounded range, e.g. below 500. From another perspective, though, comparing percentages of the two groups above or below a particular value can be seen as paying attention to aggregate features of *all* of the data since, paraphrasing one student speaking about the Amherst-Holyoke data set, 'knowing that 55%…are above the cut point means that 45% are *below* the cut point.' A comparison of *counts* above a cut point across two distributions, however, does not take all of the data into account except in the rare case of equal sized groups—an important distinction that we will discuss in more detail below.

There are several ways that cut point reasoning can go awry, however. When cut points distinguish only a small portion of a data set, just a few points on one side of the cut point and the rest on the other, they essentially serve to identify unusual values or outliers and conclusions based on these subgroups may not be robust. In addition, cut points only describe an aggregate view of data when comparing the percentage of data on either side of a *single* cut point. We contrast this use of a single cut point with a form of pair-wise comparison of values between two groups that uses "slices" of a distribution, i.e. portions of a distribution formed by *more than one* cut point, so that the distribution is divided into more than two sections. Performing a comparison using several slices is, indeed, using a classifier view of data. The difference between "slices" and the divisions formed by a single cut point is that a single slice in the middle of a distribution effectively disregards everything else about the distribution—information that could radically change an interpretation based solely on the data in the slice itself. Several examples of these problems with slices follow.

Comparing slices

Although we have no examples of teachers exploring the AIDS data using pair-wise comparisons across internal slices, we have seen both teachers and students make such comparisons using the

Amherst/Holyoke data set. These observations mirror those that other researchers (Konold & Pollatsek, 2002; Watson & Moritz, 1999) have noticed in students comparing groups.

For example, when two teachers were using the Amherst/ Holyoke data to answer the question "In which town do students spend more hours doing homework per week?" they proceeded to divide the data into seven bins. They then agreed to discount both the bottom bin (less than four hours) and the top four bins (greater than 12 hours) to focus just on the two "middle" bins in which most of the data were clustered (see Figure 4). They argued that they had "clear explanations" for what was going on with both the bottom and top groups, so that it was acceptable to exclude them from the analysis. Specifically, they argued that the number of students doing fewer than four hours of homework a week was the same in both schools and could be discounted because it would add nothing to the comparison. By contrast, they said, "The top is a lifestyle no matter where they live," and therefore should also be discounted because it didn't represent something about the typical student. Here is another example of teachers' preferring binned data (as in Figure 2) to fully separated data (as in Figure 3) so that they could understand the data in easily visible chunks, even though they often just accepted the chunks that the software provided.

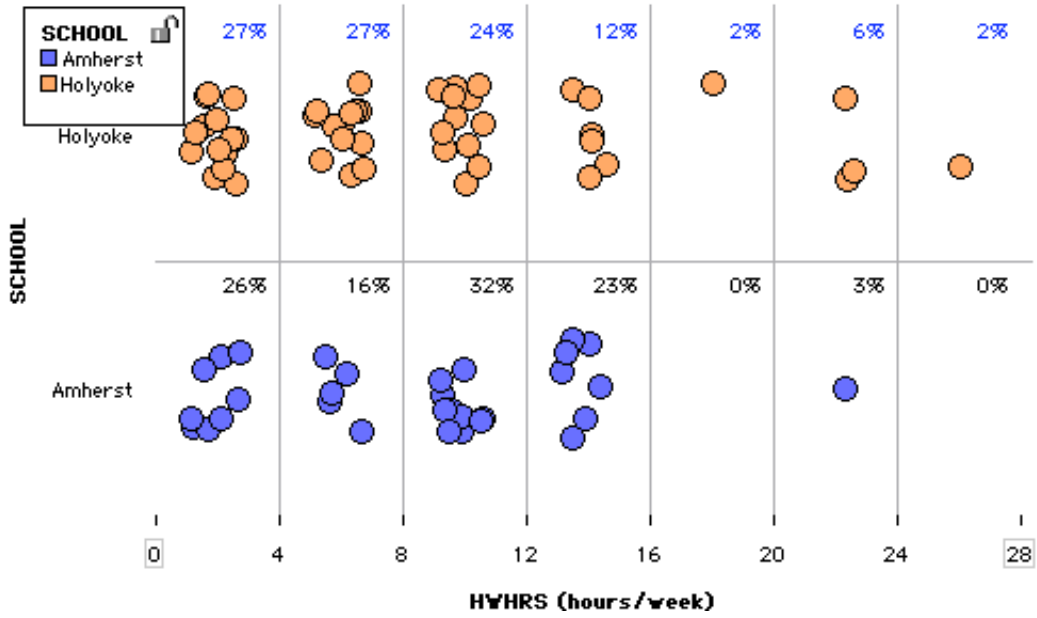

Figure 4. Looking at a slice of Homework hours in two Schools. Teachers focused on students who did between four and twelve hours of homework each week.

Having constructed this graph and chosen to focus on just the "typical" students, the teachers compared the percentage of students from each school in each of the two bins they found interesting. In the 4- to 8-hour bin, they found 27% of Holyoke students and just 16% of Amherst students; in the 8- to 12-hour bin, they found 24% of Holyoke students and 32% of Amherst students. These two comparisons within slices led them to conclude that Amherst students did more homework than Holyoke students. In fact, they were comparing the comparisons, noticing that a higher percentage of Holyoke students were in the 4- to 8-hour bin, and a higher percentage of Amherst students were in the higher, 8- to 12-hour bin. They reasoned that, since Amherst had a higher percentage of students who studied eight to 12 hours and Holyoke had a higher percentage studying between four and eight hours, Amherst students must do more homework. The teachers also looked at the data another way, noticing that the ratios of students in the two bins within each of the schools was different—roughly equal numbers in Holyoke study "8 to 12" and "4 to 8" hours respectively, whereas twice as many Amherst students study "8 to 12" hours as study "4 to 8" hours. This confirmed their view that Amherst students studied more than Holyoke students

However, there are problems with their argument. A slice-wise comparison across groups when there are multiple slices effectively ignores the rest of the distribution, i.e., if you only know about the

percentage of students in each school who study between four and eight hours, you can't really say much about the overall pattern. As stated, the conclusion doesn't account for differences in study habits of those who study more than 12 hours who, if included, might lead to a different analysis and conclusion.

We can view the teachers' approach to this analysis from the perspective of reducing variability. By putting the data into bins, they reduced the overall variability to just seven categories. By using both data-based and contextual arguments to discount the relevance of several of these bins, they further reduced the variability in the data and, in the end, compared just four numbers. This level of detail was sufficient for them until they were asked *how much more* Amherst students studied than Holyoke students. They then found their representation and analysis insufficient to answer the question. In fact, in trying to come up with some kind of number, these teachers wondered whether the horizontal position of points in each bin had any quantitative meaning, and had to be reminded that bins served merely as categories without distinguishing among their members.

Dividing distributions into multiple bins, then making an argument to disregard those on the extremes, was a common approach throughout the seminar, both using TinkerPlots™ and on paper. It is, we conjecture, an extension of the strategy of "disregarding outliers." In fact, one of the pieces of "statistical knowledge" with which several of the teachers entered our seminar was: "If you're using the mean, it's important to disregard the outliers." Or, differently stated, "If there are outliers, use the median." With a TinkerPlots™ binned representation, it is simple to disregard entire groups of "outliers" because there is no visible difference in value among points in the same bin.

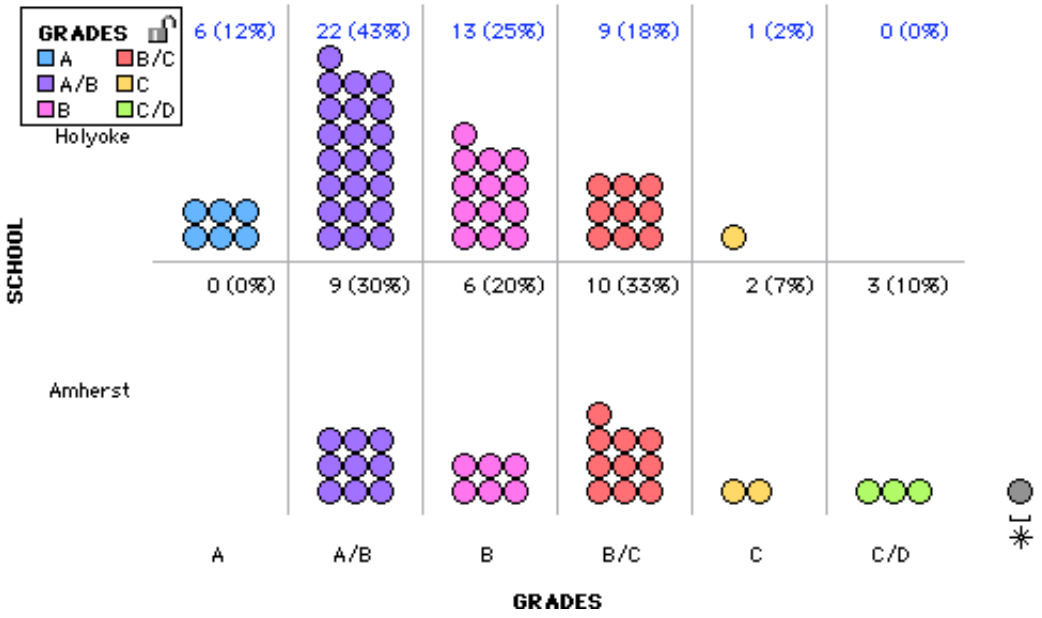

Figure 5. Comparing Grades in Schools by looking at the A/B slice

Another example of people looking at slices to compare distributions comes from the 7th grade class of one of our participating teachers. Using the same data set, but looking at grades received by students across the two schools, one student looked only at the students who received a mix of As and Bs (the A/B column, second from the left in Figure 5) to argue that more Holyoke students got better grades than Amherst students, although we don't know if she was referring to numbers or percents as they're both displayed and both point in the same direction. Likely, this student was focusing on the highest category about which she could make a comparison, since no Amherst students got straight As, though such a focus on a single slice ignores the variability and shape of the distribution.

In contrast to focusing on an internal slice created by more than two cut points, some students used a *single* cut point to split the data set between the A/B and B categories so they could compare the percentage of students in each school getting either As or As and Bs (55% Holyoke, 30% Amherst). They argued persuasively from that observation that Holyoke students got better grades. However, others thought such a comparison wasn't fair, that "you shouldn't pay attention just to the

smart kids," and wanted to look at the students getting Bs or Bs and Cs. They seemed to be torn between a desire to look at *more* of the data, expanding their view to include kids more in the middle *in addition* to the "smart kids," and wanting to *limit* their view by looking at a single slice, such as just those students getting Bs.

Another student in this discussion proposed expanding the number of students being considered by moving the cut point to just below the B range. That would mean that 80% of Holyoke students would be included and 50% of Amherst students, which seemed both like a large enough fraction of the students in the sample, and a large enough percentage difference to be able to draw the conclusion that Holyoke students get better grades than Amherst students. One could argue, although these students didn't, that setting the cut point below B is really more of a comparison of the lower end of the distribution than of the upper end, i.e., which students get worse grades? Again, we see a tension in students' techniques between, on the one hand, *narrowing* their perspective on the data by using bins to reduce the variability and number of points they have to attend to and, on the other hand, *expanding* the scope of data they're considering to include a minimum number of students with which they feel comfortable.

The ease of creating bins in TinkerPlots™ supported both cut point representations and the possibility of focusing just on internal bins as "slices," which produced a kind of argument that we conjecture would not have occurred as often otherwise. For some middle school students, the distinction between slices and cut points remained problematic. In general, the distinction between slices and cut points presented fewer problems to the teachers in our seminar after the beginning, but one teacher continued to routinely disregard the ends of a distribution for contextual reasons and focus on the middle. For example, in analyzing the weight of student backpacks before and after a hypothetical assembly on the topic of back problems caused by heavy backpacks, it was this teacher who disregarded those carrying less than 4 pounds as being "slackards."

4.2. PERCENTAGES AND PROPORTIONAL REASONING

Working with unequal size groups brings up the issue of additive versus multiplicative reasoning. Figure 6 illustrates the difference between additive and multiplicative reasoning. In this binned representation, both numbers and percents are displayed in each bin. Thus, it is possible, and even made relatively simple by TinkerPlots™, to compare the *number* of subjects with T-cell counts above 500 in the Standard vs. Experimental subgroups (using additive reasoning). In the case of Figure 6, this would lead to the incorrect conclusion that the Standard protocol was more effective because there are more subjects above 500 in the Standard condition. Of course, the correct way to make this comparison is by using percents, using multiplicative reasoning. Several of the teachers in our seminar correctly compared the two drug protocols by looking at the *percentage* of patients with T-cell counts above the "healthy" cutoff of 500. Interestingly, however, several teachers in our seminar struggled with the distinction between an analysis based on numbers of points and one based on relative percents. The tendency to use counts rather than ratios was surprisingly robust.

Another task we gave teachers was to judge if the Experimental treatment was as good for women as it was for men. Figure 6 is one representation that could support this analysis. In each bin, every man is colored green and every woman is colored red. By visually separating the males and females in each bin, teachers looked at the rough proportion of men to women in each of several ranges of T-cell counts. A further step one group of teachers took was to create a circular fuse in each bin to create pie graphs (Figure 7). These two representations, though, provide very different views of the data, since in Figure 7, the salient part of the representation is the ratio of men to women (green to red) and the number of cases in each bin is displayed only as a number. In Figure 6, however, we *see* which bins have more data, but the pattern of proportions is less apparent. Note that Figures 6 and 7 have different bin boundaries than Figures 2 and 3; depending on the sequence of steps the user has taken to arrive at the graph, bin boundaries may be placed slightly differently. The user can also specify the number of bins and the bin boundaries.

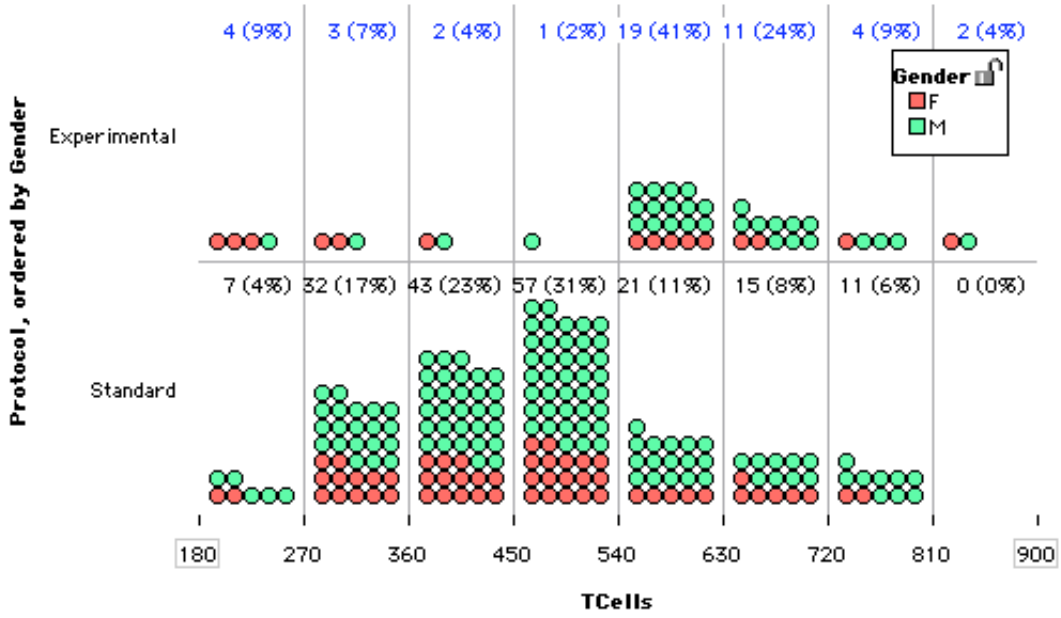

Figure 6. Binned representation showing proportions and numbers simultaneously

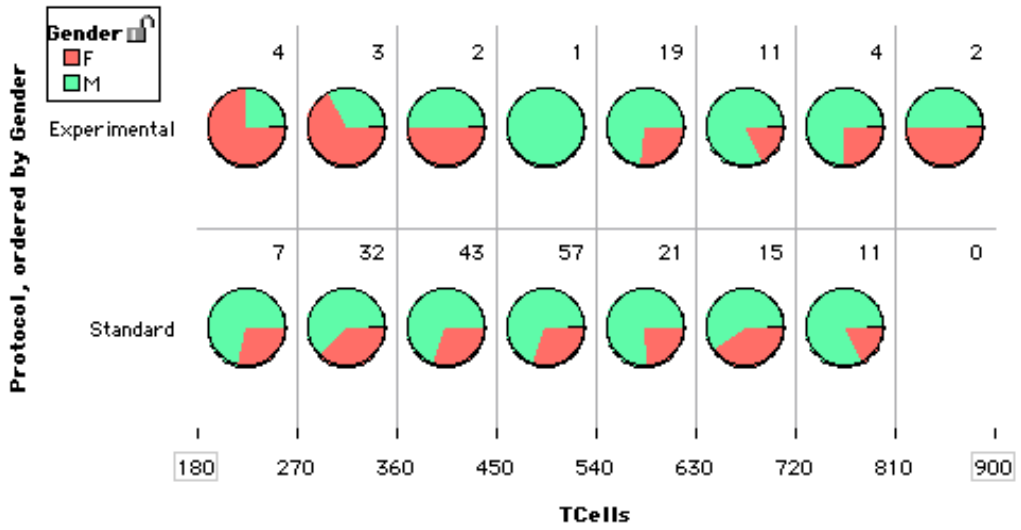

Figure 7. Pie graphs showing shift in gender proportions for Experimental protocol and rough consistency for Standard protocol

Teachers made an unusual argument using Figure 7. Some teachers used these pie graphs to notice patterns in proportions across the entire data set, specifically the rough shift of the gender proportions from low to high T-cell counts for the Experimental protocol (more females had low T-cell counts; more males had high T-cell counts) compared with the rough consistency of the gender distribution for the Standard protocol. The pie graphs enabled teachers to emphasize the fraction of the data of each gender in each bin precisely because the numbers were visually normalized into a single circular whole for each bin, even though the numbers in each bin were different. For several teachers, this was a most compelling graph because of the salience of the visual pattern. One teacher said, "I like the circle because it captures the whole….I can see the three-fourths for that region much better on a pie chart than with a histogram."

Even though the counts are displayed in each bin, the pie graph representation makes it difficult to discern information about numbers in each bin, numbers that must be big enough to draw conclusions with any confidence. For example, there is only one person in the 450 – 540 range for the Experimental protocol, so the fact that the circle is all green gets more visual "weight" than it should.

One teacher found this problem quite disconcerting. She said, "You may like it, but I don't. I think it distorts the numbers." Instead, she preferred Figure 6, which she could visually inspect for proportionality, while also being able to see the relative numbers going into that proportion. "I prefer the histogram because I can actually see the counts. With pie charts you think percent[age]s, you don't think numbers." Still, Figure 7 supports a kind of argument that is not salient in Figure 6 and it is an argument that bears consideration.

While this particular graph (Figure 7) was compelling for a subset of the teachers, pie charts seldom appeared after this data set. It is possible that the teachers who didn't like the pie charts because they "distorted the numbers" convinced the others that, despite its simplicity, this graph was likely to be misleading.

As noted above, some learners are troubled by how these representations hide absolute numbers, while others are not. We've seen both students and adults look at sets of pie graphs and forget that some of them represent only three points while others represent 30. When this is pointed out to them, they *do* know that it would be easy to dramatically change the proportions in the set of three by changing only one data point, and therefore they're less likely to trust that proportion than the proportion in a pie graph representing more data points. Still, we have seen many teachers forget about this concern unless it comes up in conversation. Although we do not have enough evidence to know why, we hypothesize two contributing factors. First, several teachers described how easily they saw patterns in a sequence of pie graphs; visually these patterns are much more striking than the numbers in the bins. Second, we also know that at the time the group worked on the AIDS data, their appreciation of the effects of sample size was relatively weak, so they may not have focused on that aspect of the representation.

Using proportions in the form of percentages or pie graphs to equalize groups of unequal size is a powerful and sometimes new idea for students in middle school. In the 6th grade teaching experiment conducted by one of the authors, several students were excited when they realized that by using percentages to compare groups of different sizes, they could "make it seem like they're even." This was much preferable to other ideas they had been considering to deal with unequal group sizes, primarily removing points from the larger group until it had the same number as the smaller (Watson & Moritz, 1999). Students were uncomfortable with this solution mostly because they couldn't figure out which points to cut without introducing a bias.

An interesting related issue arose in the 6th grade group among some students who weren't wholly comfortable with proportional reasoning. When students were using percentages even though they knew that the groups were different sizes, some worried that each "percent" *meant a different thing* in each of these groups—that is, ten percentage points may have been six students in one group, and eight students in the other group. How could they compare percentages when they meant different numbers of students? These hardly seemed equivalent. A similar issue arose in a discussion with a VISOR teacher discussing the money earned each week by a sample of Australian high school students. "It's very confusing because if you're realizing four girls equals seven percent whereas only one boy equals four percent…I mean if I didn't put the numbers, I could have just said, 'Okay, percentage-wise the boys make more [money per week].' But if you look really at how many kids each of those really affect..." In all these examples, we see a tension for both students and teachers between recognizing the power in being able to "make groups even" by putting everything on a common scale of 100, and distrusting that transformation and being drawn back to worrying about absolute numbers.

In these examples, we see teachers and students using proportional representations to deal with the variability of group size that is often encountered in data. We also see them struggling with how to simultaneously retain information about group size which is often put in the background, if not completely hidden, when emphasizing proportional views. TinkerPlots™ provides ways to represent both proportional and count information, with relatively more emphasis on one or the other in different views. Pie graphs, for example, focus on proportions without considering absolute counts. By making all these combinations of representations possible—counts without percents, or percents without counts in bins—TinkerPlots™ provides a wide choice of representations and possible

arguments. This, in turn, forces students and teachers to confront and discuss the conclusions that can be drawn, legitimately or not, from each representation.

4.3. BINNING IN THE CONTEXT OF COVARIATION

The teacher seminar had gone on for several months before we approached covariation. By that time, the strategies described above—binning of various sorts, using cut points, comparing slices, using both additive and multiplicative reasoning—had been thoroughly explored, but only in the context of comparing groups in which a single numerical distribution is partitioned into bins, and compared across one categorical attribute. But how would binning and the relationship between counts and percents play out in the context of covariation where there are two numerical variables?

Interestingly, teachers often extrapolated their binning methods to work in a 2-dimensional covariation situation, taking advantage of TinkerPlots™' tools to easily create bins on each axis, thereby partitioning the plane into a "matrix" of boxes (similar to Konold, 2002). For example, in an analysis of a data set on states, relating percent who voted for Bush in the 2000 election to median age, some teachers produced a graph of this "matrix" form by creating four bins on the X-axis and five on the Y-axis for a total of 20 boxes (see Figure 8). Each cell in the matrix contains a "rectangular fuse" of all the data points that belong there. Each small square in a fused rectangle of data points represents a single point—a state—and its color represents the Bush vote; darker colors represent larger percents, as illustrated in the legend in the top right. The data points have not been ordered within each rectangle, so the colors create a checkerboard pattern.

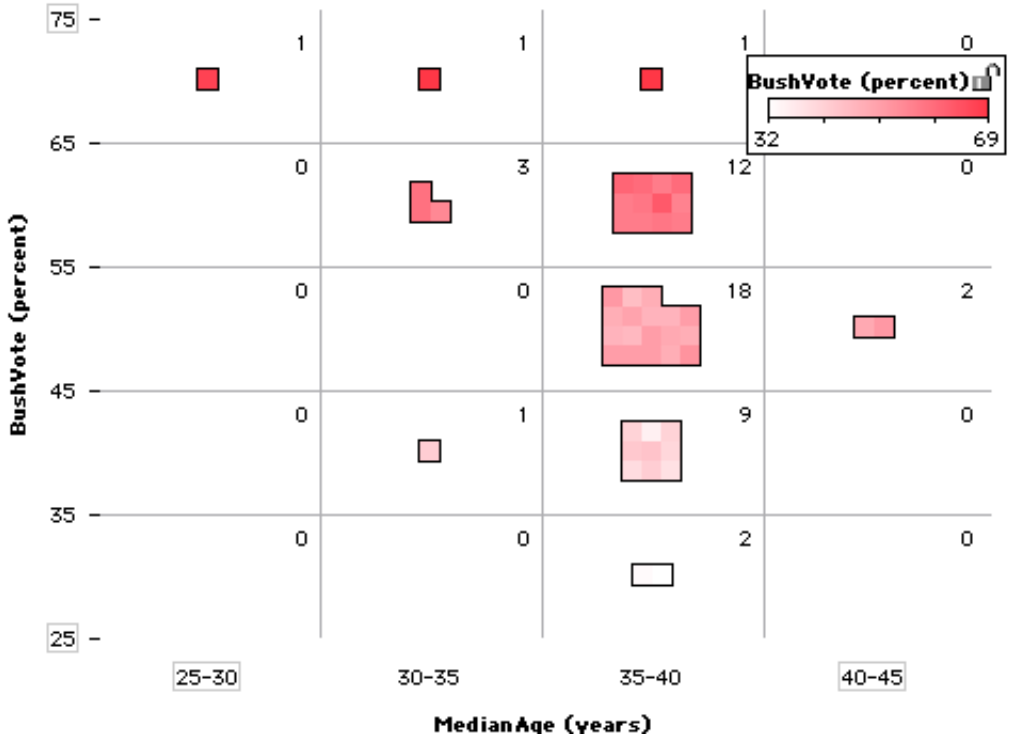

Figure 8. "Matrix" of Median age by Percentage voting for Bush in 2000

Here, the X-axis (median age) is divided into four bins, each representing a 5-year interval. The Y-axis is divided into five bins, each representing a 10% interval of votes for Bush. Thus, for example, the box that includes states whose median age is between 35 and 40 AND which voted between 45% and 55% for Bush has 18 states in it.

Characterizing relationships in these kinds of data is still difficult, even after reducing the variability by binning in this way, and the statements teachers made based on these graphs reflected that. One of the most interesting ways teachers described this kind of graph was to make non-

probabilistic statements of the form, "No country with a life expectancy under 40 has a female literacy rate over 50%." This kind of statement essentially takes the "stochastic" out of statistics and reduces variability still further by finding an "absolute" statement that one might imagine being true for other samples. Using Figure 8, teachers made statements like: "Any state in which the median age is between 40 and 45 has a Bush vote between 45% and 55%," or, "Any state in which the Bush vote is between 25% and 35% has a median age of 35 to 40." In making these statements, teachers were noticing columns or rows of the matrix that have only one cell filled; if more than one cell in a column or row is filled (e.g., the row of states in which 35% to 45% voted for Bush), these kinds of statements can't be made. Note that while these statements are strictly true according to the graph, they are each based on only two states and ignore much of the variability in the data. They are not reflections of the underlying process that may link these two variables, a process that is inevitably noisy and that won't produce exactly the same results every time (Konold & Pollatsek, 2002).

Teachers often migrated towards descriptions that focused on the deterministic, non-stochastic statements they could make from the data rather than the noisy processes underlying the bigger picture. In fact, we've noticed that some people seem to actively avoid making statements about the likelihood of certain features of data being due to chance, preferring instead to handle variability by finding subsets of the data about which they can make more deterministic claims. People want to say: "If I do X, Y will happen. If that's not always true, I can try again with a more precise understanding of the relationship between X and Y." For some people, such a belief in a microscopic version of causality is preferable to the necessity of confronting an inherently variable process. Coming up with a story that predicts exactly some, even if not all, of a data set removes those items from consideration in a stochastic framework.

We might consider the teachers' focus on a small subset of the data in this kind of deterministic way as an example of a "classifier" view of the data (Konold et al., 2003) since the teachers appear to be attending to a small set of data points without considering their relationship to the rest of the distribution. While their thinking does have some "classifier" characteristics, these teachers are thinking in a more complex way. They have *some* awareness of the rest of the data set, since the cell being described must be picked out from, and is therefore seen in relation to, the other cells in that row or column. Still, like cut points that isolate only a few unusual points, this kind of a view doesn't consider much of the data at once. Using Figure 8 to create an argument of this kind does not create an overall description of the relationships between the variables.

There are other examples, however, of a covariation "matrix" in which this kind of argument would be more defensible. For example, using the same States data set, one of the teachers produced the graph shown in Figure 9. In describing this graph, this teacher called attention to "the empty quadrant" in the upper right, which enabled him to say something like, "If a state spends at least $6789 per student, its teen birth rate will be less than 46 per 100,000." The teacher created this graph by placing horizontal and vertical lines at the means for each variable and then felt comfortable making the statement even though there are actually three states where educational spending is above $6789 but whose teen birth rate is more than 46 per 100,000. We conjecture that it was the *form* of the argument he was concerned with, more than the exact details; he knew that there were values of Ed_Spending and TeenBirths for which a statement like his would actually be true.

In fact, other teachers who made similar graphs used slightly different lines in order to completely isolate the "empty quadrant" so that it contained NO points. Is the teacher who left in a few points more comfortable with variability than those who excluded all points before drawing conclusions? Note that while this is a deterministic statement similar to the one above relating median age to percentage voting for Bush, it takes into account characteristics of the entire data set, since there is a significant visual "hole" in the graph, a more global than local phenomenon. And, while this teacher did not explicitly describe these data as representing a signal and noise, one can imagine his statement turning into an appropriate stochastic claim.

It is interesting to note that Figure 9 does in two-dimensions what Figure 3 seems to do in one-dimension. It retains a detailed display of the overall shape of the data while marking and focusing on important or interesting regions. That is, both graphs involve a fully separated view of continuous variables and use dividers set at a contextually relevant value (in Figure 3, a T-cell count of 500), or

reference lines set at both mean values, and then online "pencil" marks (Figure 9) to point to or get information about the graph and thus, the data. While comparable information can be obtained using bins, among our teachers, use of pointing tools on top of a fully separated display is more consistently connected to global statements about the data than are binned representations. A display of the full shape of the data seems to lead to a somewhat more holistic, aggregate view; or perhaps it just doesn't *also* support a classifier view. Yet, these types of uses of tools are less common and perhaps more difficult to conceptualize than are methods of binning. It's not clear whether the use of certain tools enables more sophisticated thinking, or whether teachers and students who have more sophisticated ways of thinking use specific tools to express their ideas.

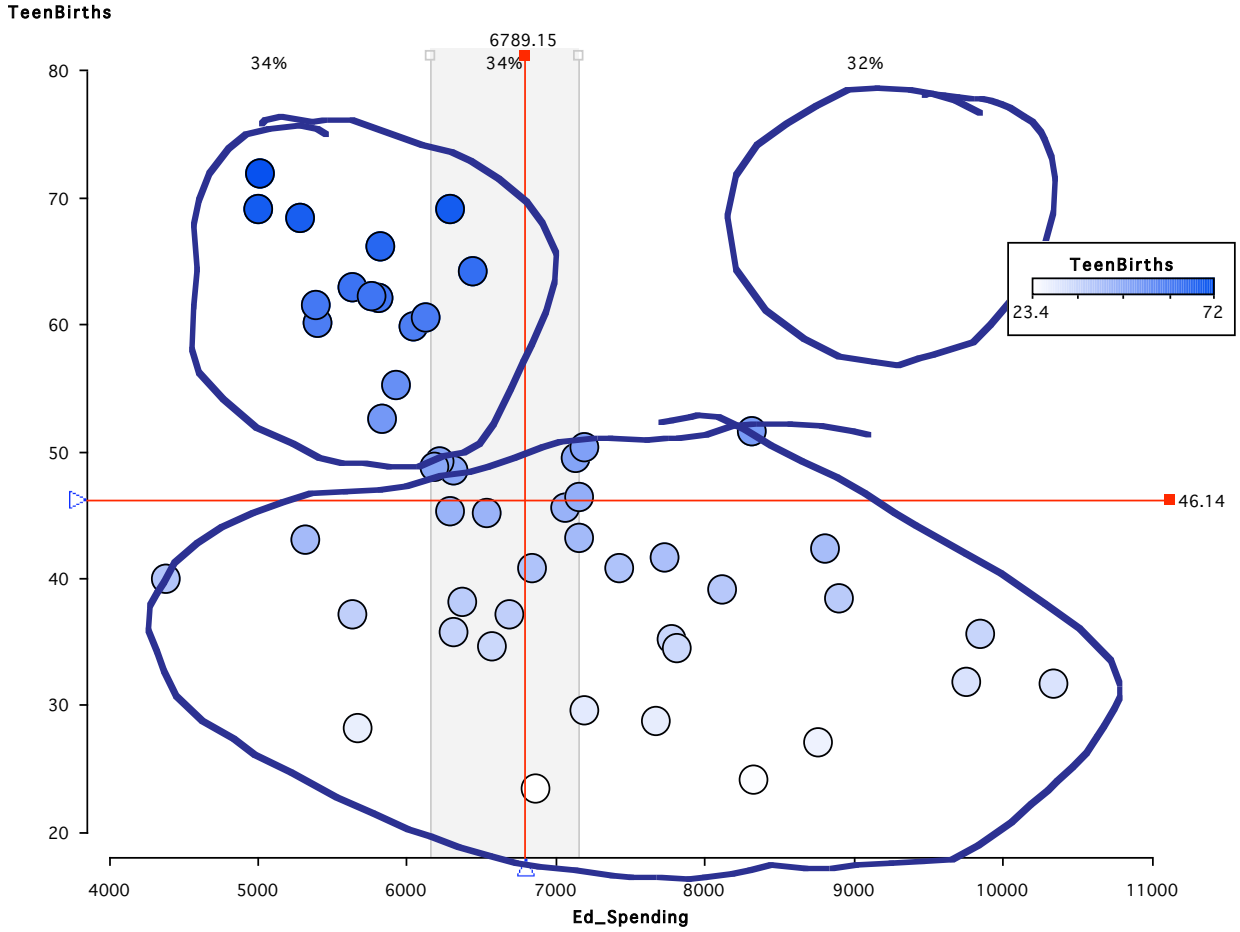

Figure 9. Finding an "empty quadrant" in the data (the circles are the teacher's)

5. CONCLUSIONS AND IMPLICATIONS

This study was designed to examine the strategies that teachers and students use to handle variability in data distributions, and to explore the possibilities for data representation and analysis that new statistical visualization tools make available. Analyzing data with such tools, in this case TinkerPlots™, makes visible reasoning that we could not have observed before. Much as new media both support the emergence of new ways of creating art and may reveal more of the artist's process, these new tools enable new representations and in so doing give us a window into the reasoning behind them. The tools offer the opportunity to examine reasoning strategies that build on the new representations they afford, as well as provide the cauldron within which these strategies emerge.

Our observations of teachers dealing with variability in the context of comparing groups agree with those of Konold et al. (1997): using measures of center was by far less common than the other strategies we have described in this paper. Our experience is, in fact, that *seeing* a distribution makes

it harder to accept a measure of center, especially a mean, as being representative of the entire distribution, with its particular spread and shape. In this sense, the binning capacity of TinkerPlots™, we believe, filled a "need" these teachers had to describe a distribution when the variability was perhaps more compelling than any measure of center. So our focus on binning was both because it was a new and very visible function of TinkerPlots™ and because we saw teachers using it in ways that helped solve a data analysis problem for them. Thus, we saw teachers using binning from early on and as time went on, they continued to use bins but became more sophisticated in their use of them. They became able to indicate more specifically which bin lines would help them make their points or to use the more flexible, but more complex, dividers to make their argument for a difference between groups (e.g. see Figure 3, in which teachers created a divider at a T-cell count of 500).

The difficulty of describing portions of distributions using percents rather than counts, i.e., multiplicatively rather than additively, is one that has been documented by other research (Cobb, 1999), and we know that students and teachers struggle with this distinction using paper versions of dot plots. But two features in TinkerPlots™ highlight the issue. First, the presence of bins in comparing groups of unequal size provides an immediate need for multiplicative thought — how does one compare a bin in one distribution with the same bin in the other? How does one interpret the changing numbers as bins grow or shrink? Second, because TinkerPlots™ can provide the count and/or the percent of values in a bin, i.e., calculating the percent is no more difficult than counting the points, both values are equally available and the conversation about appropriate measures is quite likely to come up, as it did in our examples. We also note that looking at percents and NOT counts (as in Figure 7, in which there was a pie chart in each bin) can also lead one astray. After several months, most of the teachers reasoned multiplicatively most, but not all, of the time. Surprisingly, however, some teachers continued to slip back into the additive type of reasoning, although they could be "pulled back" out of it through conversation.

How do these results relate to Konold's taxonomy described above (Konold & Higgins, 2002; Konold et al., 2003)? The two relevant pieces of that taxonomy in this context are "classifier" and "aggregate." Those using a "classifier" view tended to view slices of the distributions out of context of the rest of the distribution, e.g. by comparing the number of students who make Bs between Amherst and Holyoke as a way to compare the entire two distributions. Even looking at these two bins as percents rather than counts doesn't solve the problem. Because the analysis ignores the distributions on either side of the bin of interest, it counts as a "classifier" view and does not answer questions about the entire distribution. On the other hand, making statements that take the entire distribution into account (e.g. 25% of this distribution is above 500 and 75% of the other distribution is above 500) is an example of "aggregate" thinking. So is using a measure of center such as the mean, but because of how easy it was for teachers to make other kinds of graphs, they rarely used measures of center.

Since TinkerPlots™ makes classifier and aggregate comparisons equally easy, the stage is set in a classroom for discussions of alternate analyses. Valuable perspectives can emerge from questions such as: Do two arguments about the same data set that rely on different representations always agree? What might be the relationship between an argument based on a classifier view and one based on an aggregate view of the same data set? Comparing data sets can also lead to new ideas: In what ways are the bins in the Amherst/Holyoke data similar to those some teachers created in analyzing the AIDS data? In what ways are they different? Because it is also possible to display measures of center and variability (e.g. box plots) on a graph, classroom discussions can include comparisons of using bins, cut points, and/or aggregate measures as a basis for analysis.

Teaching with a tool like TinkerPlots™ requires an in-depth understanding of the kinds of thinking the tool might engender and make visible. Once thinking is made visible, it can be discussed, challenged, and made more robust. But becoming aware of student thinking also raises new challenges for teachers, as these new ideas can be difficult to comprehend, their validity can be difficult to assess, and helping students work from them to more canonical ideas involves navigating complex and unexplored conceptual terrain. It is our experience that there is no substitute in teaching with a new tool for using it first as a learner, especially in a group in which alternate views of data and representation are likely to arise, as they certainly will in the classroom. Using a tool as a learner can

also help teachers experience the importance of complex data sets that pose particular challenges, e.g., unequal group sizes, and encourage the asking of more than one question.

We realize that many of our hypotheses and tentative conclusions are not "statistical" in nature, but we believe that the kind of study that follows a small group of teachers or students using new tools over two years can uncover new ways of thinking that a shorter and more controlled study could not. The teaching experiment model upon which this research is based (Ball, 2000; Cobb, 2000; Steffe and Thompson, 2000) uses a different set of methodologies and has different goals from a study that can be evaluated with a quantitative statistical analysis. A teaching experiment is based on a two-phase cycle that includes a development phase and a research phase. This cycle occurs multiple times during any teaching experiment, roughly after every session, in the course of planning for the next one. Teaching experiments make it possible to gather rich, multi-dimensional data that takes into account the interactions of individual students, the classroom social context, the resources available and the teacher's facilitation. A teaching experiment can identify multiple learning paths, as the methodology takes as a basic principle that individuals have different learning "stories."

The flip side of a teaching experiment's power to look deeply at a complex progression of events is that the information it provides is based on a small number of participants, e.g., a classroom of students, who are not guaranteed to be representative of a larger group. Based on our research, there are three categories of generalization that we feel are worthy of future study. 1) Generalizing to other groups. Do teachers not in the special subset of "those who were willing to collaborate with us" use TinkerPlots™ in different ways? How do middle school students use TinkerPlots™ to deal with variability? What difference would it make if we had just middle school teachers or just high school teachers in our seminar? 2) Generalizing to other interactive visualization tools. There are a few similar tools on the market now (TableTop™, Fathom™) and there are certain to be more. What aspects of these tools have similar affordances to TinkerPlots™? In what ways do they support different approaches to variability? 3) Generalizing to other professional development situations. We worked with this group of teachers for two years, and they were willing to stick it out for that long. What would happen in a shorter course? Do teachers use TinkerPlots™ differently if they have access to the software at home, not just in school? In addition to these issues of generalization, we have observed that some learners were more able to describe and discuss data as a sample. We have some preliminary evidence that this awareness may affect the representations that learners create and find compelling, but that is just the beginning of another complex story.

Our study explored the thinking of teachers and students as they grappled with, and tried to make intelligent use of, new software tools. An exploratory study of this kind, which involved researchers both as educators and observers of complex behaviors in a new arena, is bound to raise more questions than it answers, and we believe that is part of its value. We hope that our conclusions, while based on selective and at times incomplete evidence, can provide researchers and teachers with new ideas as well as with new research hypotheses regarding the role of new software tools in statistics education.

ACKNOWLEDGEMENTS

Support for this paper comes from the National Science Foundation (NSF) under grant REC-0106654, Visualizing Statistical Relationships (VISOR) at TERC, Cambridge, MA. The views expressed are those of the authors and do not necessarily reflect those of the NSF. A version of this paper was presented at The Third International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-3), Lincoln, Nebraska, USA, in July 2003. The authors are listed in alphabetical order; both contributed equally to the document.

Thanks to the teachers and students who have so graciously shared with us their thinking about these often difficult ideas. Thanks to our colleagues at TERC and SRTL and to the anonymous reviewers from SERJ whose questions and critiques have helped us improve upon earlier drafts of this paper. This work could not have been done at all without the contributions of Cliff Konold, designer of TinkerPlots™ and Bill Finzer, designer of Fathom™. Iddo Gal's comments as SERJ editor were

invaluable in helping us to continually sharpen our argument. And a special thanks goes to our incredibly adept research assistant, Camilla Campbell.

REFERENCES

Bakker, A. (2004). Design research in statistics education: On symbolizing and computer tools. Doctoral dissertation, Utrecht University.

Bakker, A., & Gravemeijer, K. P. E. (2004). Learning to reason about distribution. In J. B. Garfield & D. Ben-Zvi (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 147–168) Dordrecht, Netherlands: Kluwer Academic Publishers.

Ball, D. L. (1991). Research on teaching mathematics: Making subject matter knowledge part of the equation. In J. Brophy (Ed.), *Advances in research on teaching* (Vol. II, pp. 1–48). Greenwich, CT: JAI Press.

Ball, D. L. (1993). With an eye on the mathematical horizon: Dilemmas of teaching elementary school mathematics. *Elementary School Journal, 93*(4), 373–397.

Ball, D. L. (2000). Working on the inside: Using one's own practice as a site for studying teaching and learning. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 365–402). Mahwah, NJ: Lawrence Erlbaum Associates.

Ball, D. L. (2001). Teaching with respect to mathematics and students. In T. Wood, B. S. Nelson & J. Warfield (Eds.), *Beyond classical pedagogy: Teaching elementary school mathematics* (pp. 11–22). Mahwah, NJ: Lawrence Erlbaum Associates.

Ball, D. L., Hill, H. C., Rowan, B., & Schilling, S. G. (2002). *Measuring teachers' content knowledge for teaching: Elementary mathematics release items, 2002*. Ann Arbor, MI: Study of Instructional Improvement.

Borko, H., & Putnam, R. T. (1995). Expanding a teacher's knowledge base: A cognitive psychological perspective on professional development. In T. R. Guskey & M. Huberman (Eds.), *Professional development in education: New paradigms & practices* (pp. 35–65). New York: Teachers College Press.

Case, R. (1978). Implications of developmental psychology for the design of effective instruction. In A. M. Lesgold, J. W. Pelligrino, S. D. Fokkema & R. Glaser (Eds.), *Cognitive psychology and instruction* (pp. 441–463). New York: Plenum Press.

Case, R. (1980). Intellectual development and instruction: A neo-Piagetian view. In A. E. Lawson (Ed.), *1980 AETS yearbook: The psychology of teaching for thinking and creativity,* (pp. 59–102). Columbus, OH: ERIC Clearinghouse for Science, Mathematics, and Environmental Education.

Clement, J. (2000). Analysis of clinical interviews: Foundations and model viability. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 547–589). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Cobb, P. (2000). Conducting classroom teaching experiments in collaboration with teachers. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 307–333). Mahwah, NJ: Lawrence Erlbaum Associates.

Cobb, P., Gravemeijer, K. P. E., Doorman, M., & Bowers, J. (1999). Computer Mini-tools for exploratory data analysis (Version Prototype). Nashville, TN: Vanderbilt University.

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

Duckworth, E. (1996). "*The having of wonderful ideas*" *and other essays on teaching and learning* (2nd ed.). New York: Teachers College Press.

Hancock, C. (1995). *Tabletop*. Cambridge, MA: TERC/ Brøderbund Software.

Hancock, C., Kaput, J. J., & Goldsmith, L. T. (1992). Authentic enquiry with data: Critical barriers to classroom implementation. *Educational Psychologist, 27*(3), 337–364.

Harel, G., & Confrey, J. (Eds.). (1994). *The development of multiplicative reasoning in the learning of mathematics*. Albany, NY: SUNY Press.

Hill, H. C., & Ball, D. L. (2003). *Learning mathematics for teaching: Results from California's Mathematics Professional Development Institutes*. Ann Arbor, MI: University of Michigan.

Key Curriculum Press. (2000). Fathom™ Dynamic Statistics™ Software (Version 1.0). Emeryville, CA: Key Curriculum Press.

Konold, C. (2002). Alternatives to scatterplots. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics*, *Cape Town, South Africa*. [CDROM] Voorburg, The Netherlands: International Statistical Institute.

Konold, C., & Higgins, T. L. (2002). Highlights of related research. In S. J. Russell, D. Schifter & V. Bastable (Eds.), *Developing mathematical ideas (DMI): Working with data casebook,* (pp. 165–201). Parsippany, NY: Dale Seymour Publications.

Konold, C., Higgins, T. L., Russell, S. J., & Khalil, K. (2003). Data seen through different lenses. Unpublished manuscript, Amherst, MA.

Konold, C., & Miller, C. (2004). TinkerPlots™ Dynamic Data Exploration (Version Beta 1.0). Emeryville, CA: Key Curriculum Press.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education, 33*(4), 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: Proceedings of the 1996 IASE Round Table Conference.* 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. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics*, *Cape Town, South Africa*. [CDROM] Voorburg, The Netherlands: International Statistical Institute.

Lamon, S. J. (1994). Ratio and proportion: Cognitive foundations in unitizing and norming. In G. Harel & J. Confrey (Eds.), *The development of multiplicative reasoning in the learning of mathematics,* (pp. 89–120). Albany, NY: SUNY Press.

Meletiou, M., & Lee, C. (2002). Student understanding of histograms: A stumbling stone to the development of intuitions about variation. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics*, *Cape Town, South Africa*. [CDROM] Voorburg, The Netherlands: International Statistical Institute.

Miles, M. B., & Huberman, M. (1984). *Qualitative data analysis.* Beverly Hills, CA: Sage Publications.

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

Mooney, E. S. (2002). A framework for characterizing middle school students' statistical thinking. *Mathematical Thinking and Learning, 4*(1), 23–63.

Mooney, E. S., Hofbauer, P. S., Langrall, C. W., & Johnson, Y. A. (2001). Refining a framework on middle school students' statistical thinking. Paper presented at the 23rd Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, Snowbird, UT.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage.

Rubin, A. (1991). Using computers in teaching statistical analysis: A double-edged sword. Unpublished manuscript, Cambridge, MA.

Rubin, A. (2002). Interactive visualizations of statistical relationships: What do we gain? In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics*, *Cape Town, South Africa*. [CDROM] Voorburg, The Netherlands: International Statistical Institute.

Rubin, A., & Bruce, B. (1991). Using computers to support students' understanding of statistical inference. *New England Mathematics Journal, 13*(2).

Rubin, A., Bruce, B., & Tenney, Y. (1990). Learning about sampling: Trouble at the core of statistics. Paper presented at the Annual Meeting of the American Educational Research Association.

Rubin, A., & Rosebery, A. S. (1988). Teachers' misunderstandings in statistical reasoning: Evidence from a field test of innovative materials. In A. Hawkins (Ed.), *Proceedings of the ISI Round Table Conference, Training Teachers to Teach Statistics*, Budapest, Hungary.

Russell, S. J., Schifter, D., Bastable, V., with Higgins, T. L., Lester, J. B., & Konold, C. (2002). *DMI Working with data: Casebook & facilitator's guide*. Parsippany, NJ: Dale Seymour Publications.

Russell, S. J., Schifter, D., Bastable, V., Yaffee, L., Lester, J. B., & Cohen, S. (1995). Learning mathematics while teaching. In B. S. Nelson (Ed.), *Inquiry and the development of teaching: Issues in the transformation of mathematics teaching* (pp. 9–16). Newton, MA: Center for the Development of Teaching, Education Development Center.

Saldanha, L., & Thompson, P. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*, 257–270.

Schifter, D. (1997). *Learning mathematics for teaching: Lessons in/from the domain of fractions*. Newton, MA: Center for the Development of Teaching at Education Development Center, Inc.

Sedlmeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making, 10*, 33–51.

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

Sowder, J. T., Philipp, R. A., Armstrong, B. E., & Schappelle, B. P. (1998). *Middle-grade teachers' mathematical knowledge and its relationship to instruction: A research monograph*. Albany, NY: SUNY Press.

Steffe, L. P., & Thompson, P. (2000). Teaching experiment methodology: Underlying principles and essential elements. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 267–306). Mahwah, NJ: Lawrence Erlbaum Associates.

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*(1), 44–70.

JAMES K. HAMMERMAN & ANDEE RUBIN
TERC
2067 Massachusetts Ave.
Cambridge, MA, 02140
USA