SAYING THE SAME (OR A DIFFERENT) THING: HOW SHAPE AFFECTS IDEAS ABOUT DISTRIBUTION IN A SOFTWARE EXPLORATION ENVIRONMENT

James K. L. Hammerman and Andee Rubin
TERC, United States
jim_hammerman@terc.edu

*Educational software for statistics and data analysis provides a variety of tools for seeing and expressing ideas about data distributions. However, the ideas that learners find important to express often depend on an interaction between software and the shape of the distributions themselves. In this interview study of teachers participating in the VISOR professional development program, we investigate how distributional shape (symmetric or skewed) and choice of software tool (*TinkerPlot *or* Fathom*) affect how teachers discuss data distributions when comparing groups. We find teachers' confidence is increased when different measures or ways of viewing data "say the same thing," which more often holds true with symmetric distributions. When these seem to conflict, typically with skew distributions, teachers work to understand the measures themselves, and introduce new ways of characterizing data, so that they can make coherent sense of the distributions. The paper introduces a distinction between* rule-driven *and* value-driven *measures which we find important in understanding teachers' analytic methods.*

INTRODUCTION

As people try to make sense of data, their approaches and conclusions grow out of an interaction among the depth and breadth of their own understanding of data, characteristics of the data distributions themselves, and the affordances provided by representational tools for displaying and manipulating data. People come to data with particular purposes, conceptual models, and initial ideas about what to expect. They use tools—including tables, charts, graphs, spreadsheets, statistical techniques, and interactive software—to try to represent these initial ideas and models, to gather information about the data and, sometimes, about the population from which they are drawn. In the process, they begin to draw conclusions about the meaning of the data, but they also change and refine their models, and then seek ways to represent these new ideas. That is, the models shape what people do with the tools—how they represent the data, the relationships they choose to explore—but the tools, in turn, shape the models of data that people hold. It is this reciprocal relationship between people's ideas about data and the representations they use to express those ideas that we explore in this study.

Exploration of the interactions between people's ideas and representational tools has been important generally in mathematics education. Noss and Hoyles (1996) state that representations provide a medium that supports enhanced communication between people, essentially adding a third "language" in which to talk about a mathematical concept. They comment that the languages available by default are natural and mathematical languages; the first "badly tuned to rigorous and precise discourses like mathematics and the latter precisely the reverse. The <representational scheme> affords a half-world in which…articulation and rigor can be made to converge" (p. 6). Other researchers have applied this perspective to a variety of mathematical concepts (e.g., Harel and Confrey, 1994; Kaput, 1994; Lehrer, Schauble, Carpenter, and Penner, 2000).

We are especially interested in how people use relatively new, *interactive* software tools such as *TinkerPlot*™ (Konold and Miller, 2004) and *Fathom*™ (Finzer, 2005) to represent and explore data. A number of prior studies describe how learners—children and adults—use a variety of such software tools to analyze data (Bakker, 2004; Bakker and Gravemeijer, 2004; Cobb, 1999; Hammerman and Rubin, 2004; Hancock, Kaput, and Goldsmith, 1992; Rosebery and Rubin, 1990; Rubin, 2002; Rubin and Hammerman, 2006). These studies highlight how ideas about data develop when people discuss them in the context of interactive tools for representing those ideas.

One theme that emerges from several of these studies is the importance and difficulty of developing an aggregate view of the data, rather than seeing them as a collection of individual cases (Bakker and Gravemeijer, 2004; Cobb, 1999; Hammerman and Rubin, 2004; Hancock, Kaput, and Goldsmith, 1992; Lehrer and Schauble, 2004). Konold, Higgins, Russell, and Khalil

(2003) describe a framework of increasing levels of complexity in how people understand data, arguing that children see data in *several* simpler ways—what they call *pointer* (focusing on the data collection event), *case* (focusing on single cases), and *classifier* (focusing on groups of cases) perspectives—before ever noticing *aggregate* (focusing on the whole data set) and emergent features of data. Because adults rarely if ever use pointer or case perspectives, and when they do their thinking is not limited to these views as it is for children, we focus on teachers' uses of classifier and aggregate views and how software supports representations of these perspectives.

CONTEXT AND METHODS

In this paper, we describe the results of an interview study in which a small group of middle and high school teachers explored two different shaped distributions using the innovative exploratory data analysis tools, *TinkerPlot*™ and *Fathom*™. Our research was conducted during the late spring of 2004, as part of the Visualizing Statistical Relationships (VISOR) project at TERC, in Cambridge, Massachusetts, USA. The nine VISOR teacher-participants met for three hours, biweekly after school at TERC throughout the 2003–04 academic year to explore a variety of data sets, and to discuss ideas about data analysis and statistics, about students' thinking about data, and about teaching. They received a stipend for their participation in the seminar, and the interview was considered part of the seminar.

The six high school and three middle school teachers (four women and five men) varied in their statistical expertise and prior experience with data analysis, from those who had been teaching AP Statistics for many years and were familiar with software data analysis tools, to those with limited experience teaching about data or using software. They taught in various Boston area schools, including a low-income, urban, mostly Latino/a district; several mixed race and SES urban districts; a mostly white, working-class suburb; and a mostly white, middle-class suburb.

The study involved intensive 45–90 minute clinical interviews (Clement, 2000) of teachers analyzing two attributes of a single data set using the teacher's choice of *TinkerPlot* (six teachers) or *Fathom* (three teachers). The data were from a survey of South Australian teenagers concerning some personal attributes (gender, age, height), and how students spend their money and time (Konold and Miller, 2004). Teachers worked with samples of 60 students from this larger survey. To explore how distributional shape would influence how teachers analyze data, we focused on the frequency distributions of two variables: height, which was distributed symmetrically, and money earned per week, which was a skewed distribution. The data set also included information on each student's gender, birth year, and year in school, and teachers focused on relationships between these and height or money earned. Figures 1 and 2 use *TinkerPlot* to show each of the focal distributions separated by gender, with the mean (Δ) and median (⊥) displayed, and dividers delineating roughly the middle 50% of the data. Although teachers also explored relationships with Year in school, we will not describe those here.

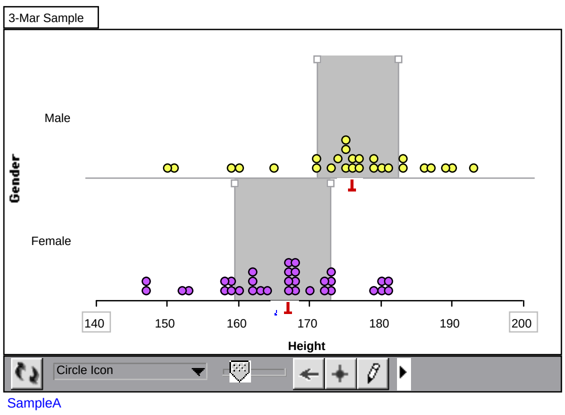

Figure 1: Height data by Gender:
Mean, median and IQR marked

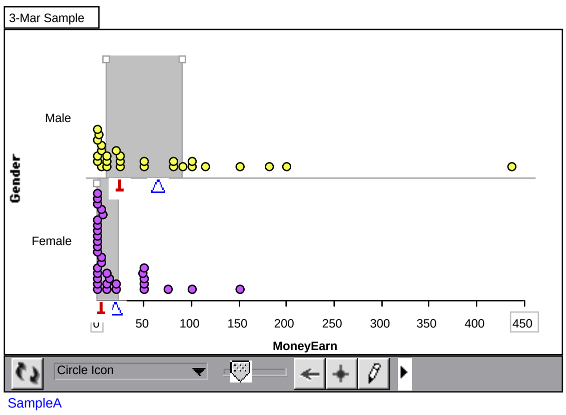

Figure 2: Money Earned data by Gender:
Mean, median and IQR marked

Interviews were videotaped and audiotaped and teachers' work on the computer was recorded using a video-feed. We transcribed the audiotapes and coded the text using categories both from our theoretical perspective and those arising from the data themselves (Glaser and Strauss, 1967). We used the video-feed to recreate the graphs teachers made.

RESULTS

Although there was much of interest in these interviews, we focus here on how teachers explored the different shaped distributions using the two software tools, and what this says about their understanding of data and statistical group comparisons. In the process, we describe a new conceptual distinction—between *rule-driven* and *value-driven* comparisons—that helps us categorize some of the ways teachers use software tools to analyze data (see also, Rubin, Hammerman, Puttick, and Campbell, 2005). While there were similarities and differences between the affordances provided by the two tools—e.g., both offered ways to display the mean or median; both offered box plots but *TinkerPlot* also provided moveable dividers and so, teachers never used box plots with *TinkerPlot*—these differences will not be the focus of our analysis.

Many standard statistical techniques involve *rule-driven* comparisons. These use a rule to calculate a value, usually a measure of center such as a mean or median, for each of the distributions being compared. The locations of these values are then compared to determine if one is “significantly” different from the other(s). Determining the significance of a difference involves other considerations, such as the variability and size of the data sets, but the basic structure of a rule-driven comparison is to find corresponding values in each data set and compare their locations. Because an analyst must accept a single value as sufficiently characterizing a distribution that it can be used “alone” to make a comparison, this approach usually fits Konold *et al.*’s (2003) “aggregate” category (although not always—a rule comparing the location of the “maximum value” of two distributions could be case-based reasoning).

In contrast, a *value-driven* comparison begins with a single value and looks at how much of each data set lies on either side of that value. Visually, this amounts to drawing a line—often called a “cut point”—through all of the distributions being compared and looking at the number or proportion of each dataset on either side of the line. The cut point value can be contextually relevant (e.g., a passing grade on a standardized test), driven by the shape of the data (e.g., a salient gap), or even a rule-based measure such as the mean or median of one group. Whether this approach shows an aggregate or classifier view depends on a variety of inter-related factors: e.g., a) the number of cut points used (an internal “slice” lying between two cut points is more likely to be a classifier view; other configurations are aggregate or unclear), b) whether comparisons are made with absolute numbers or proportions of points (proportions are more likely an aggregate view; numbers are less clear), c) the size of the subgroups on either side of the cut point (relatively large groups are more likely an aggregate view), and other factors (Hammerman and Rubin, 2004; Hammerman, Rubin, Puttick, and Campbell, 2005; Rubin and Hammerman, 2006).

We found teachers using both rule-driven and value-driven comparisons with both *TinkerPlot* (TP) and *Fathom* (F), although value-driven comparisons were more common with the money earned (ME) data than with height (H) data. We believe that teachers find rule-driven measures of center (e.g., mean and median) a more convincing description of the symmetric height data than the skewed money earned data because they are near one another and seem to point towards the “modal clump” (Konold *et al*., 2002). In this way, they seem to “*say the same thing*”—both the mean and median of boys’ height is about 9 cm. larger than that of girls. With the skewed, money earned data, rule-driven measures of center are further from one another and teachers didn’t feel they described the typical data very well. Teachers in our study responded to this situation in several ways: 1) They struggled to understand what these different measures were telling them. 2) They tried to make the skewed data more symmetric by filtering out the zero earners and an extreme high value, usually arguing that they were only interested in earnings of those with jobs. 3) They more often looked for non-standard, value-based comparisons that could capture salient characteristics of *these* data—e.g., the preponderance of girls among the zero earners, or differences in the number or percentage of each gender among high earners (those above \$50/ \$60/ \$70/ \$80). When these measures, along with rule-driven measures, all pointed to boys earning more money than girls, teachers gained confidence (Hammerman and Rubin, 2004).

Rule-driven comparisons were common with both software tools. In both data sets, teachers using either tool compared the positions of the male and female mean and median. For example, “Daniel” (a pseudonym, like all teacher names herein) (*TinkerPlot* (TP), money earned (ME)) used gender differences in both means and medians to conclude there was a real difference in money earned. “*For the males, the mean is like 51 whereas for the females it's like 20. And for the males the median is like 24, whereas the females it's like 5. I don't think we need any [more] quantification than that to get the EEOC [Equal Employment Opportunity Commission] involved.*” David (*Fathom* (F), height (H)) used means to conclude, “*Boys are on average 9 centimeters taller than girls*” and also compared by quartiles. “*So the first quartile is the same for boys and girls. This is with the filter* [to temporarily exclude zero earners] *on, and the third quartile is very different for boys and girls. For girls it's around 50, for boys it's around 100.*”

Not all rule-driven comparisons were of this canonical type, though. One teacher, Robert (TP, H) consistently looked at the mean and median in bins, identifying a range of values in which they lay rather than a single point, and counting numbers of each group above the overall mean to make a kind of value-driven comparison. Late in the interview, though, when asked how much taller boys were than girls, he used the relative location of these bins to make a rule-based comparison of heights. “*I can say that the boys' mean and median is this number of centimeters* [175 to 179] *and the girls is 165 to 169, so the difference is 10 centimeters taller than the girls*.”

Use of value-driven comparisons was more complicated, and also more common with the money earned data set than with the height data. A few teachers used specific values they thought were interesting to make comparisons. For example, Sharon (TP, ME) compared the numbers of males and females above a cut point of AU$70 per week that marked a gap in the distribution. “*There's only 9 percent of the female population for this sample is above 70. In the boys, it's 16...[counting] 36. So 9 versus 36 percent is above 70. But now I'm looking over here, and I'm seeing that 70 percent of that female population, 70 percent of the sample are in the 0 to 13 range, and 42 percent of the males.…So there's a lot of females making little money, and there's a lot more males making more money.*” Similarly, Natalia (TP, ME) used differences in the numbers and proportions of high earners—defined as those making over AU$50—to characterize differences in earnings at different grades.

Several teachers, using both *TinkerPlot* and *Fathom* compared the numbers of each gender making no money, essentially using zero as a cut-point value. For example, Alice (TP, ME) said, “*And there are 15* [females earning zero] *versus males, 4. So the fact that the females are not earning, or have a lower earning average is because a lot of them don't have jobs, I assume, or they're not having paid jobs. So that's important.*” Patty (F, ME) drew a similar conclusion, saying, “*It's interesting that there are more females making nothing, not having a job at all, not making any money, than the males.*”

With both *Fathom* and *TinkerPlot*, many teachers used values from rule-driven measures to make their comparisons. Robert’s approach was interesting. He used the bin-based mean and median as cut points in a value-driven comparison. For example, to compare groups by gender, he counted the numbers of boys and girls above the bin in which the combined mean was located. “*If you look at all the yellows* [males], *all these kids, all the boys, geez only five girls are taller than the mean or median. All the others are boys. And you got a lot of purple* [females] *down here.*” (TP, H)

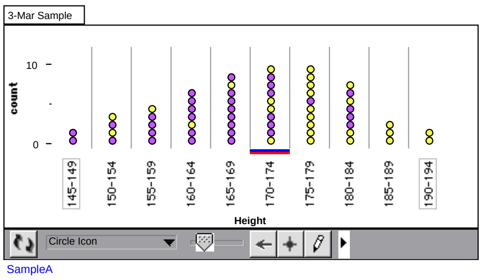

Figure 3: Counting numbers above the mean bin

More typically, teachers used a rule-driven measure from one group as a cut-point, and counted the number or percentage of the other or both groups above that point. The simplest example (David, F, H) involved counting the number of boys taller than the tallest girl, where the rule “tallest girl” finds a specific height, and then comparison is made by the numbers of boys (7) and girls (by definition, zero) who are taller than that. In another example, Alice (TP, ME) said, “*42 percent of the males earn less than 20* [the female mean], *and if I compare the females now…We've got 35 percent of the males above their* [own, male] *mean and only 9 percent of the* *females. And, you know, basically 90 percent of the females earn less than the males* [that is, the male mean]. *So that's a pretty big difference for me.*"

A still more complex merging of rule-driven and value-driven comparisons involves looking for values that are common (or close) for two different rule-based measures (a task which is only sometimes possible), and then comparing groups based on these measures of distributional shape. For example, using *Fathom*'s built-in box plot feature, David (F, ME) compared the first quartile of males with the median of females. "*That first quartile for boys goes up to 5, which is where the median...it's the same as the median for the girls*." Here, David is using the median and first quartile to decide that AU$5 is an important value; but his comparison uses that value as a cut point and describes the percentage of each distribution (50% of girls v. 25% of boys) who earn less than that. By contrast, a rule-based comparison would look at the relative location of the median, or of the first quartile, for both groups, but wouldn't compare the two across groups.

Gary, who often characterized the broad shape of the data by describing groups that are above or below particular values, did something similar to David. Also using box plots, Gary (F, ME) compared the median of males with the third quartile of females: "*But still half the males make more than three quarters of the females, which is still pretty interesting.*" Daniel (TP, H) manually adjusted the dividers in *TinkerPlot* to mark and then compare the middle 50% (IQR) of heights. "*The middle half of females tends to fall below the middle half of males…So that's, the first quartile in the males is 173 centimeters, and the [third] quartile in the females is 172.*" In essence, all three of these observations are value-based comparisons—e.g., that 50% of girls, but only 25% of boys make less than AU$5; that half the boys make more than three fourths of the girls; or that 75% of boys are taller than 75% of girls. Daniel's statement is an observation about the position of the "modal clump," defined here as the interquartile range (IQR) and so, could also be considered a rule-driven comparison.

DISCUSSION AND CONCLUSION

The shape of the data affects how people analyze it. As noted above, symmetric data usually yield rule-driven measures of center that are close together, which increases people's confidence in them. Rule-driven measures of center in skewed data are in different places, and people respond in a variety of ways, including by creating value-driven measures which can be more easily adjusted to fit and describe specific features of data sets and which may be easier to understand. When various measures yield the same *conclusion* in several different ways, that increases teachers' confidence.

As we show here, the concept of rule-driven and value-driven comparisons can be a powerful tool for understanding how learners—here, teachers, but also, students—analyze data using software tools. We encourage continued work to elaborate and refine this schema, including descriptions of the different types of measures learners invent and their characteristics, and how learners use measures in different circumstances. Analysis of how different software tools support the creation of different ways of looking at data is hinted at here and would also be important.

Teachers in this study also sought to make inferences beyond these data to the larger population of Australian teenagers, often by looking at consistency of findings across several samples, which they saw as another example of data "saying the same thing." In doing so, they were informally invoking the logic of t-tests (rule-driven measures) and chi-square tests (value-driven measures). Our current work continues to explore these issues of data comparison and the development of inferential thinking in software data exploration environments.

ACKNOWLEDGMENTS

Research reported in this paper was funded by the United States' National Science Foundation, Grant #REC-0106654. The opinions expressed are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

Bakker, A. (2004). *Design Research in Statistics Education: On Symbolizing and Computer Tools.* Unpublished Dissertation, Utrecht University, Utrecht, The Netherlands.

Bakker, A. and Gravemeijer, K. P. E. (2004). Learning to reason about distribution. In J. B. Garfield and D. Ben-Zvi (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Clement, J. (2000). Analysis of clinical interviews: Foundations and model viability. In A. E. Kelly and 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.

Finzer, W. (2005). *Fathom*™ Dynamic Data™ Software (Version 2.0). Emeryville, CA: Key Curriculum Press.

Glaser, B. G. and Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New York: Aldine Publishing.

Hammerman, J. K. and Rubin, A. (2004). Strategies for managing statistical complexity with new software tools. *Statistics Education Research Journal,* 3(2), 17-41.

Hammerman, J. K., Rubin, A., Puttick, G. M., and Campbell, C. (2005). Developing models of distributions using *TinkerPlot.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, PQ.

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

Harel, G. and Confrey, J. (Eds.). (1994). *The Development of Multiplicative Reasoning in the Learning of Mathematics*. Albany, NY: SUNY Press.

Kaput, J. J. (1994). The representational roles of technology in connecting mathematics with authentic experience. In R. Biehler, R. W. Scholz, R. Strasser and B. Winkdelmann (Eds.), *Didactics of Mathematics as a Scientific Discipline*, (pp. 379-397). Dordrecht, Netherlands: Kluwer.

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

Konold, C., and Miller, C. (2004). *TinkerPlot*™ Dynamic Data Exploration (Version 1.0). Emeryville, CA: Key Curriculum Press.

Konold, C., Robinson, A., Khalil, K., Pollatsek, A., Well, A., Wing, R., *et al*. (2002). Students' use of modal clumps to summarize data. In B. Philips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics (ICOTS-6), Cape Town*. Voorburg, The Netherlands: International Statistical Institute.

Lehrer, R., and Schauble, L. (2004). Modeling natural variation through distribution. *American Educational Research Journal,* 41(3), 635-679.

Lehrer, R., Schauble, L., Carpenter, S., and Penner, D. (2000). The interrelated development of inscriptions and conceptual understanding. In P. Cobb, E. Yackel and K. McClain (Eds.), *Symbolizing and Communicating in Mathematics Classrooms: Perspectives on Discourse, Tools, and Instructional Design*, (pp. 325-360). Mahwah, NJ: Lawrence Erlbaum Associates.

Noss, R. and Hoyles, C. (1996). *Windows on Mathematical Meanings: Learning Cultures and Computers*. Dordrecht, Netherlands: Kluwer.

Rosebery, A. S. and Rubin, A. (1990). Teaching statistical reasoning with computers. *Teaching Statistics,* 12(2), 38-42.

Rubin, A. (2002). Interactive visualizations of statistical relationships: What do we gain? In B. Philips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics (ICOTS-6), Cape Town.* Voorburg, The Netherlands: International Statistical Institute.

Rubin, A. and Hammerman, J. K. (2006). Understanding data through new software representations. In G. Burrill (Ed.), *Thinking and Reasoning with Data and Chance—2006 NCTM Yearbook.* Reston, VA: National Council of Teachers of Mathematics.

Rubin, A., Hammerman, J. K., Puttick, G. M., and Campbell, C. (2005). The effect of distributional shape on group comparison strategies. Paper presented at the The Fourth International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-4), Auckland, New Zealand.