DISTRIBUTION: A RESOURCE FOR UNDERSTANDING ERROR AND NATURAL VARIATION

Rich Lehrer and Leona Schauble
University of Wisconsin-Madison
USA

The studies we present investigate elementary students' reasoning about distributions in two contexts: (a) measurement and (b) naturally occurring variation. We first summarize an investigation in which fourth-graders measured the heights of a variety of objects and phenomena, including the school's flagpole, a pencil, and several launches of model rockets. Students noted that the measurements were distributed and that sources of error corresponded to differences in qualities of distribution, especially spread. Next, students investigated the distributions of measurements of height for rockets of different design, to learn whether and how they could be confident that rockets with rounded nose cones "really" went higher than those with pointed nose cones. We then turn to the naturally-occurring variation context, in which these same students (now fifth-graders) studied the growth of Wisconsin Fast Plants™, fast-growing members of the Brassica family that enable multiple cycles of classroom observation and experiment within a school year (life cycle is about 40 days). We recount how students became adept at using changing shapes of distributions to support plausible accounts of growth processes. Questions about what would be likely to happen "if we grew them again" motivated investigations of sampling, which, in turn, suggested choices of statistics to represent a sample distribution. Finally, students invented means for considering how one might know whether two different distributions of measures could reasonably be considered "really different."

INTRODUCTION

However one classifies objects and events, their variability is at least as important as their similarity, yet variability is given very short shrift in school instruction—students are given few, if any conceptual tools to reason about variability. Typically, these tools consist only of brief exposure to a few statistics (e.g., for calculating the mean or standard deviation). In contrast, we focus on data modeling (Lehrer & Romberg, 1996). Students typically participate in contexts that encourage them to develop questions, consider qualities of measures and attributes relevant to their question, and then go on to structure data and make inferences. Within the scope of data modeling, distribution affords an organizing conceptual structure for thinking about variability,

Accordingly, we conducted a teaching study that spanned two years with a single intact class and teacher. We began in the spring school semester with fourth-graders (much of this work was conducted by our colleague, Anthony Petrosino; cited in Petrosino, Lehrer, & Schauble, in press) and then the following spring semester, continued with many of the same students. Both studies encompassed approximately eight weeks of daily instruction lasting an hour to an hour and a half, and concluded at the end of the semester with individual interviews of participating students. The fourth grade study involved a series of tasks and tools aimed at helping students consider error as distributed and as potentially arising from multiple sources. Distribution was structured in this context as an account of the "true" value of clusters of measurements and their associated error. In the fifth grade, we switched our focus to naturally occurring variation. Students studied the growth of Wisconsin Fast Plants™ and characterized growth as change in the shape of distributions of measurements. Thus, the studies follow the historical evolution of ideas about distribution (Konold & Pollatsek, in press; Porter, 1986; Stigler, 1986).

METHOD

The teacher, Mark Rohlfing, worked with us to engage in coordinated cycles of (1) planning and implementing instruction and (2) studying the forms of student thinking—including both barriers and resources—that emerged in the classroom. These two forms of activity were concurrent and coordinated; the development of student thinking was conducted in the context of particular forms of instruction, and instructional planning was continually informed by and locally contingent upon what we were learning about student thinking. Twenty-two students (12

boys, 10 girls) participated as part of a school-wide reform initiative for teaching and learning of mathematics and science (Lehrer & Schauble, 2000). The teacher had 11 years of teaching experience at the time of this study, but neither the students nor the teacher had previous experience with the forms of mathematics and science that were pursued. The major sources of data were field notes, classroom videotapes, the inscriptions of data that students created during the study, and transcriptions of the concluding interviews.

DISTRIBUTION: A RESOURCE FOR UNDERSTANDING ERROR

Classroom Instruction: First, we review the series of tasks that we designed to help students consider error as distributed and as potentially arising from multiple sources. In the context of these tasks, students reasoned about potential sources and mechanisms that might produce variability in light of the structure in the variation that they observed. Sources of random error included individual differences among measurers, instrument variation, and replication variation. To distinguish between random and systematic variation, students conducted experiments on the design of rockets (rounded vs. pointed nose cones). The purpose was to determine whether differences between the resulting distributions of rocket heights (obtained with round vs. pointed nose cones) were consistent with random variation, or if the differences could be attributed to the design of the rockets.

To make characteristics of center and spread of distribution salient, instruction began with students measuring the height of their school's flagpole using a hand-made, paper "height-o-meter." We expected that students would readily agree that the flagpole had a "true" height and therefore, the center of the distribution might be interpreted as its approximate measure. Second, the make-shift nature of the instruments might inspire ideas about why one might expect some measurements above and below the actual height of the flagpole. Symmetry in measure could then serve as a potential explanation for symmetries in the shape of the distribution of measurements. Third, we hoped it would occur to students to group similar values of measurements, thus introducing the important mathematical idea of a distribution as the function of the density of the number of observations within particular intervals.

The measurements that students recorded ranged from 6.2 m to 15.5 m, with a median value of 9.7 m. Mr. R challenged students to arrange the measures to show "your best sense of how you would organize this." As expected, students' displays of these data varied, providing them an opportunity to understand how different representational choices made aspects of the data more and less visible. Mr. R used these displays to provoke consideration of the idea of interval, extreme values, and the possibility that perhaps not all the measurements were equally likely. As students discussed their confidence in the measures, they began to argue that more central values were the ones that seemed most trustworthy. Most students argued that more measurements within an interval were an indication that the actual flagpole height was somewhere in that bin, and that the true height was less likely to be in bins with fewer cases. However, others argued for considering a middle region that encompassed the middle intervals. Students felt that this middle region resulted from the measurement process itself, which would create values "a little over" and "a little under," and that extreme values represented being "a lot over or under."

To accentuate relationships between measurement error and the resulting characteristics of distribution, students next used a ruler to measure a regular No. 2 pencil. When asked to compare the relative precision of the flagpole and pencil measurements, students gestured with their hands to indicate the relative compactness of the intervals in each display. Mr. R pressed students to consider *how much more precise* each measurer was in the pencil context, as compared to the flagpole context. A researcher used the metaphor of a difference as "the distance and direction you have to travel" from the center of the distribution to the value in question. Students worked in groups to calculate distances from the median for the flagpole and pencil measurements and "binned" their differences. The resulting distributions of difference scores recapitulated the sense of relative compactness observed in the original measures, with students noting that the flagpole differences were "much more spread out." The notion of "typical" spread was introduced as the median of the distribution of differences. Students went on to measure the height of the flagpole with a plastic instrument and found that their measurements were far less

variable than those collected with the paper "height-o-meter. This finding led to attempts to catalog contributions to error, including "human" error (i.e., "hand movement"), method variation, instrument precision, and qualities of the object itself (e.g. being able to grasp the pencil, but not the flagpole).

The concluding phase of instruction emphasized an experiment as a contrast between two distributions of values. Students first deployed three launches of a model rocket with a rounded nose cone. They used their plastic tools to record the apex of the rockets for each trial. These launches set the stage for considering individual difference and trial variation as components of error. Next students conducted "experimental" launches with a pointed nose cone, expecting that these would "cut through" the air and reach higher altitudes. Finally, they compared the resulting distributions to decide whether or not rockets with rounded or pointed nose cones would reliably go higher. This comparison was structured by their teacher by constructing a reference distribution of rounded nose cones, partitioned into thirds by the median and spread number. Thus, one interval included all cases below the median and spread number, the next encompassed the median with the lower and upper bounds determined by the spread number, and the third included the remaining cases above the median and spread number. Students were surprised to find that 86% of the measurements of rocket altitude obtained from pointed nose cone launches fell into the lowest interval of the reference distribution.

Assessment of Student Learning: At the completion of instruction, students completed two forms of assessment. First, all students completed a group-administered paper and pencil test consisting of two released items from the 1997 National Assessment of Educational Progress, designed for students at Grade 4. These items assessed students' capabilities to compare tables and to read and interpret pictographs. In addition, we conducted individual clinical interviews of about 30 minutes each with 15 of the students to assess strategies for reasoning about the probable goal of an experiment given the data, comparing distributions of unequal numbers, and reasoning about measurement variability. We report some of these results.

Reasoning about Experiment: The first interview item was adapted from research conducted with middle school students (Petrosino, et al., in press). Participants reviewed a table of data and attempted to infer the likely purpose of the experiment that motivated the data collection. All 15 interviewees produced a plausible response about the likely purpose of the experiment. However, they volunteered that the data were not sufficient to support a definitive conclusion. Students pointed out that the distributions presented for comparison overlapped substantially, and that there were too few data points to be confident that the distributions really differed. Several students spontaneously made reference to "what we did in class with the medians and spread numbers." Students suggested comparing the medians of the two distributions, eliminating outliers, or obtaining additional data. In Petrosino's previous research, middle school students had rated themselves "very confident" that the data in the table supported a definitive conclusion.

Measurement Precision and Variability: During instruction, students had measured the school's flagpole with both a paper measuring device and a plastic altimeter. Interviewees were asked to predict the measurements that might be obtained by a fictional class of 15 children who measured the height of a 100-ft statue, if the class measured first with one instrument and then with the other. Specifically, students were asked to construct the distributions that might result. As one would find if the paper tool were less precise, student-generated distributions were notably less variable for the pooled distribution of student-generated altimeter measurements (SD=13.9) than for the paper tool (SD=26.7). As anticipated, students' distributions were roughly symmetric.

DISTRIBUTION: A RESOURCE FOR UNDERSTANDING NATURAL VARIATION

Students posed and investigated a series of questions about the growth of Wisconsin Fast Plants™, members of the brassica family that have a life cycle of about 40 days, and thus support multiple cycles of classroom observation and experiment with a year. The investigations entailed posing fruitful questions, identifying and defining attributes that could be recorded over time (plant height, width, number of leaves, seedpods, etc.), and agreeing on common procedures for

measuring these attributes. Characterizing changes in the plants required students to compare distributions of measures (on each day of growth, measurements were taken for 63 plants).

On day 23 of the plants' growth, Mr. R collated the class's measurements of their plants and asked students to work in their groups to agree about what would serve as a "typical height" for a Fast Plant at this point in the life cycle. Although most readily accepted this proposal, a few students objected strenuously, insisting that since there were plants at each of several heights, all heights were "equally typical." Thus, when the idea of a "true measure" was not present to support ideas about center, some students had difficulties making the shift from thinking about particular cases (i.e., plants) to thinking about a value that could "stand in" for the collection (the entire set of plants). Over time, as we focused students on the "shape of the data," the class increasingly made this shift.

Discussions about the shape of the data usually started with consideration of "clumps," "holes," and typical values of the distribution, and progressed to ways of picturing "the shape of the distribution" on different days of growth. Early on, Mr. R asked students to agree on a way to represent the height data two distributions of plants to "show what's typical and also how spread out the measurements are." Students invented several ways of doing so, from simple ordered lists to case magnitude displays to hybrids of frequency graphs and stem-and-leaf displays. By swapping and interpreting their displays, students came to understand various senses of the "shape" of the data. These discussions were punctuated by debates about the forms of representation that best supported arguments about "what's typical" and "how spread out." Students summarized the shape of the data by recourse to indicators of center and spread, which were constituted as they investigated the effects of different partitions of the distribution on the resulting shape. Choices of statistics to represent a sample distribution were motivated by extended investigations of sampling. Students investigated what would be likely to happen "if we grew them again" by exploring distributions of samples of differing number and size. These sampling studies began by recording measurements of plant height on small cards, putting them into large envelopes, and drawing out samples of different size. They were eventually extended to work with a computer tool, which allowed flexible exploration and representation of the effects of varying the sampling plan. These investigations revealed some statistics were much less robust than others as indicators of typicality and variability. Understanding of shape was extended as students played "Make My Distribution," a game in which groups constructed distributions that fit specified constraints (median, range) and shapes (e.g., normal, skewed, uni- and bimodal). Eventually, students segmented distributions into quartiles and generated box plots consisting of what they called "hinges" and "doors" and compared the density of measures within comparable intervals. Students became reasonably adept at using changing shapes of distributions to support plausible accounts of growth processes, for example, why one would be likely to see a left-skewed distribution during the earliest days of growth, and why the center of the distribution would shift to the right as the plants' life cycle continued.

The semester culminated with studies of the effects of growth factors (light and fertilizer) on plant growth. Students compared distributions of plants grown under different conditions. Initially, they parsed the distributions of measures (initially, with respect to cut-points, then, with box plots) and comparing segments. For example: "Half of the high-light plants are taller than 15 mm on day 28, but only one third of the low-light plants grew that tall," or, "The upper hinge for the high-fertilizer plants is 17 mm, which is very close to 16.8, the upper hinge for the low-fertilizer plants." Finally, students considered how one might be confident that two distributions of measures could be considered "really different." They did this by comparing two distributions of plants grown under differing conditions and considering the degree of their separation in light of the variability one might reasonably expect "if we grew them again."

Assessment of Student Learning: At the conclusion of instruction, all students participated individually in a one-hour interview.

Shape, sampling. Children were presented with displays of two curves, one after the other. Both were labeled 20 mm at the left tail and 80 mm at the right tail, but the first was a normal curve and the second was skewed. The curves were described as the measurements of height taken on one day of the life cycle of a population of grass plants. For each curve in turn, students were asked, "Suppose we picked 23 of these plants by chance—we put them into a hat

and randomly picked out any 23. What would the overall shape of this new group of 23 look like?" Students were asked first to sketch the curve representing the likely shape of the sample of 23, and then to construct the sample by selecting values from a set of 1-inch, labeled with values. Over 85% of students drew curves that were similar in shape to the "parent." Moreover, approximately one third truncated the tails to indicate that there would be a smaller chance of choosing the measures least frequently represented in the distribution. A large majority of the students justified their sketch by talking about the roles of both chance and structure. When asked to construct a distribution using the supplied values, all students constructed a shape similar to that of the parent, and a third truncated the tails. Moreover, 85-90% of the participants explicitly pointed out that the shape of the sample would be likely to, but *need not necessarily* mimic the shape of the population.

Distribution as signature of growth. The next pair of items also concerned displays of a skewed curve and a normal curve. For each display in turn, students were asked, "When we grew our Fast Plants, what day or days did the distribution of plants have a shape that sort of looked like this one?" Students were also asked to "tell a story that explains why the distribution might have had that kind of shape." Over 70% of the students replied that the left-skewed curve represented the distribution of plant heights at the beginning of their life cycle. Although 20% justified this choice by appealing to empirical memory (that's the way they looked), most of the children proposed causal factors that might account for the observed shape, for example, the role of growth factors to account for variability (e.g., plants received varying amounts of light) or limits of growth as a reason for the observed shape of the curve ("They can't get any smaller than zero mm," or, "They all start out the same size."). Similarly, 94% identified the normal curve as the way the distribution looked when the plants were in the middle of their life cycle.

Comparing distributions. Students observed two stem-and-leaf displays representing distributions of "battery life" (that is, number of hours that a collection of batteries functioned before failing) for a set of batteries made "the usual way" and another set (of unequal number) made "a new way." They were asked which way of making batteries resulted in batteries that lasted longer.

Eighty percent of the students volunteered three or more characteristics of the distributions that supported their conclusion (20% mentioned only one). One quarter compared either the highest or lowest case values of the two distributions, but it was the sole justification offered by only two. The remaining children also mentioned a variety of more sophisticated justifications. These included: identifying a "clump" of values in one distribution that was not included in the other (13%), making qualitative reference to the location of "most of" the high or low values in one distribution or the other (11%), locating a value or "clump" from one distribution within the other (12%), comparing the medians of the distributions (20%), or mentioning the proportion or fraction of separation or overlap of the two distributions (7%). Two thirds of the students applied a sampling-like approach to comparing the distributions. That is, they referred to qualities of center, spread, and overlap of the two distributions to argue that the result would be likely to come out "the same way" if the test were run again or run with additional batteries. Most noted that some variation from the original results should be expected, and that it was therefore not possible to conclude with 100% certainty that the results would support the same conclusion. Regardless, given the qualities of the distributions under consideration, students felt reasonably confident that additional testing would be unlikely to yield a different result.

DISCUSSION

In the history of mathematics, it was difficult to negotiate the shift from reasoning about variability in the context of measure to mathematically conceptualizing variability that occurred naturally (Porter, 1986). Similarly, there were challenges in this shift that students needed to negotiate. First, the very idea that it was admissible to identify a typical value to stand for all the cases in a naturally varying set seemed foreign to some students. In contrast, it seemed readily evident to students in the context of measure that items necessarily have a "true height" and that that height could be estimated by finding the center of a set of measurements. However, with a set of plants, every height seemed equally "true." The key to negotiating this shift appeared to be a

dual focus on shape and on the idea of repeating the process of growth (e.g., "What would happen if we grew them again?") Students used distributions to make claims about growth processes, anchored to ideas about plausible biological mechanisms and functions (growth spurts, appearance of buds and seed pods, etc.). We found it particularly useful to encourage students to invent ways of describing and representing data, and then to recruit those representations to support arguments about biological mechanisms.

REFERENCES

American Association for the Advancement of Science (1993). *Benchmarks for science literacy.* New York: Oxford University Press.

National Research Council (1996). *National science education standards.* Washington, DC: National Academy Press.

Gould, S.J. (1996). *Full house.* New York: Three Rivers Press.

Konold, C. & Pollatsek, A. (in press). Conceptual underpinnings of averages. *Journal for Research in Mathematics Education.*

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

Lehrer, R., & Schauble, L. (2000). Modeling in mathematics and science. In R. Glaser (Ed.), *Advances in instructional psychology, Volume 5* (pp. 101-159). Mahwah, NJ: Lawrence Erlbaum Associates.

Petrosino, A.J., Lehrer, R., & Schauble, L. (in press). Structuring error and experimental variation as distribution in the fourth grade. *Mathematical Thinking and Learning.*

Porter, T.M. (1986). *The rise of statistical thinking 1820-1900*. Princeton, NJ: Princeton University Press.

Stigler, S.M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Cambridge, MA: Harvard University Press.