Introducing Students to Data Representation and Statistics

Richard Lehrer
Vanderbilt Univeristy
<rich.lehrer@vanderbilt.edu>

I describe the design and iterative implementation of a learning progression for supporting statistical reasoning as students construct data and model chance. From a disciplinary perspective, the learning trajectory is informed by the history of statistics, in which concepts of distribution and variation first arose as accounts of the structure inherent in the variability of measurements. Hence, students were introduced to variability as they repeatedly measured an attribute (most often, length), and then developed statistics as ways of describing "true" measure and precision. The design of the learning progression was guided by several related principles: (a) posing a series of tasks and situations that students perceived as problematic, thus creating a need for developing mathematical understanding as a means of resolving prospective impasses; (b) creating opportunities for developing representational fluency and meta-representational competence as constituents of conceptual development; (c) introducing statistics as invented measures of the qualities of distribution; and (d) adopting an agentive perspective for orienting student activity, according to which distribution of measures emerged as a result of the collective activity of measurer-agents. Instructional design and assessment design were developed in tandem, so that what we took as evidence for the instructional design was subjected to test as a model of assessment, resulting in revision to each. I conclude with a look at ongoing work to design an assessment system to measure students' understandings of data and statistics, and with some thoughts about prospective synergies between mathematics and science education.

The discipline of statistics originated in problems of modeling variability (Porter, 1986; Stigler, 1986). History has not changed all that much: Professional practices of statisticians invariably involve modeling variability (Wild & Pfannkuch, 1999), and as in other sciences (e.g., Giere, 1992), it is through model contest that statistical concepts become more widespread and stable (Hall, Wright, & Wieckert, 2007). Another lesson of history is of particular importance: Reasoning about variability was initially most prominently pursued in contexts of measurement error. Astronomers, for example, suggested that distances between stars were fixed, but that measurements varied, just by chance. Mathematical efforts to characterize the form and structure of chance gave rise to concepts and models still in use today, such as least squares fit.

Our research program follows in this historic tradition: Contexts of measure afford children entrée to a series of core conceptual structures or "big ideas" in the discipline and also, to the core disciplinary practice of inventing and revising models. Accordingly, I outline a design of instruction that features repeated measure for introducing students to practices and related concepts of data representation, statistics, chance, and modeling. These practices and concepts are all developed by students to account for observed variability in measurements. As I describe components of the design, I characterize some of the recurrent patterns of student reasoning that we observed during successive iterations of the design in fifth- and sixth-grade (10, 11 years of age) urban classrooms in the United States. These collectively establish a sense of "lessons learned". Our efforts to account for emerging patterns of student reasoning were accompanied by corresponding efforts to encapsulate these patterns of reasoning in the form of an assessment system, which is sketched in the second section of the paper. I conclude with some prospects for integrating

mathematics and science education via a shared interest in constructing and revising models of variability.

Designing Instruction to Support a Learning Progression

The instructional design was guided by an image of statistical reasoning as emerging from and enmeshed within a larger system of activity that we refer to as data modeling (Lehrer & Romberg, 1996; Lehrer & Schauble, in press). As Figure 1 suggests, data modeling is composed of two coupled systems of activity. The upper triangular region in the figure depicts the learning challenges and resources associated with the *design* of research. Designers confront challenges such as posing questions and identifying the nature of variables and their measures.

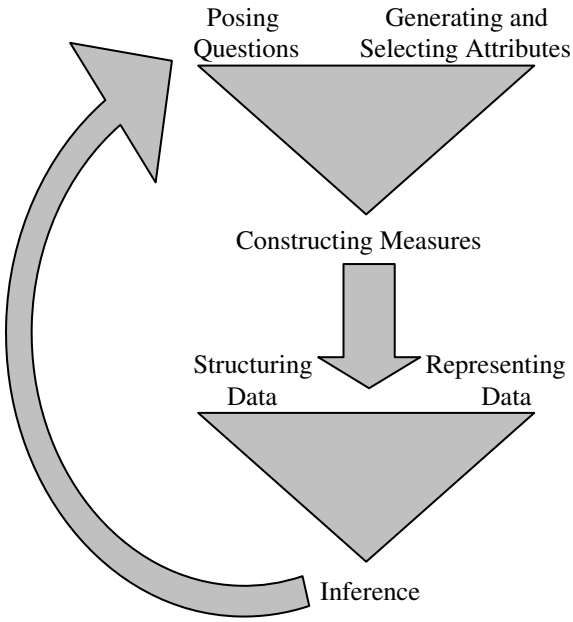

Figure 1. Schematic of data modelling.

The lower triangular region encompasses *analysis*, depicted as an interaction among data structures, representations, and models of inference. Analysts confront challenges of imposing structure on data, of choosing displays to highlight aspects of structure, and of making judgments about phenomena in light of variability and uncertainty. Although the cycle as illustrated invites inference of linear progression, in practice, these components of data modeling are typically interactive. For example, attempting to develop a measure of an attribute often profoundly alters one's conception of that attribute.

To initiate students into practices of data modeling, we designed a hypothetical learning progression – a sequence of tasks, tools, activities, and forms of argument – aimed at supporting students' development of mathematical accounts of the inherent variability of measure. The learning progression was envisioned to unfold in three coordinated phases in the classroom. In the first, students all repeatedly measured the same object and designed a representation intended to communicate trends in the collection of measurements that they noticed. In the second, students used these displays to invent statistics. One invented statistic indicated the "best guess" of the measure of the attribute of the object and the precision of the measurements. Students explored the qualities of their invented statistics

with new samples of measurements of the same object conducted with a better tool. The latter resulted in distributions that were less variable but that had approximately the same centre. The third, modeling phase included investigation by students of the behavior of chance devices and the subsequent harnessing of these devices to construct models of measurement error. In the sections that follow, I describe the rationale for each of these three phases and also suggest recurrent patterns in student reasoning that we observed as we implemented the design over several iterations in fifth- and sixth-grade classrooms in an urban school in the United States. Participating students were from under-represented groups in the United States. Their families were of lower socioeconomic status.

Inventing Representation

Students measured an attribute of a familiar object, such as the arm-span of their teacher. To measure arm-span, each student first used a 15-cm ruler and then a metre stick. Each time, students recorded the value of the measure. The aim of this initial activity was to provide students with a context in which collective properties of the data, especially distribution, could be viewed as emerging from the actions of individual agents. We anticipated that students' prior history with measurement would serve as a resource for making sense of the variability of the measurements. For example, the 15-cm ruler had to be iterated more often than did the meter stick to span the same distance. (The former resulted in greater error and hence greater variability among the measurements.)

We presented students with an unstructured collection of their measurements and challenged them to create a display (of the more variable measurements) that communicated what they noticed about the batch of data. After students created their displays, other students presented the display to the class and described what the display tended to "show and hide" about the data. This tactic was intended to foster representational fluency (Greeno & Hall, 1998). With instructor support, students compared and contrasted their invented displays. We anticipated that comparing and contrasting different displays would clarify relations between the choices made by designers and the resulting "shape" of the data. This tactic was also intended to foster meta-representational capacity (diSessa, 2004) – the capacity to view a data display as representing a trade-off. Different choices resulted in different perceptions of the shape of the same data. We were especially interested in helping students understand how displays that grouped data and counted cases within each group produced a symmetric, bell-shaped distribution. Students considered possible reasons for the bell-shape of grouped data in light of the process of measure. We concluded this phase of instruction by soliciting students' conjectures about what might happen if "we measured again".

Recurrent Patterns of Representation

The most striking feature of the displays generated by the students was their variability. Despite years of education emphasizing conventional graphs, students often found this task challenging and even daunting.

Highlighting order. The most common solution to the problem of display was to structure the data by ordering the magnitude of the cases. Some solutions were lists, such as that displayed in Figure 2.

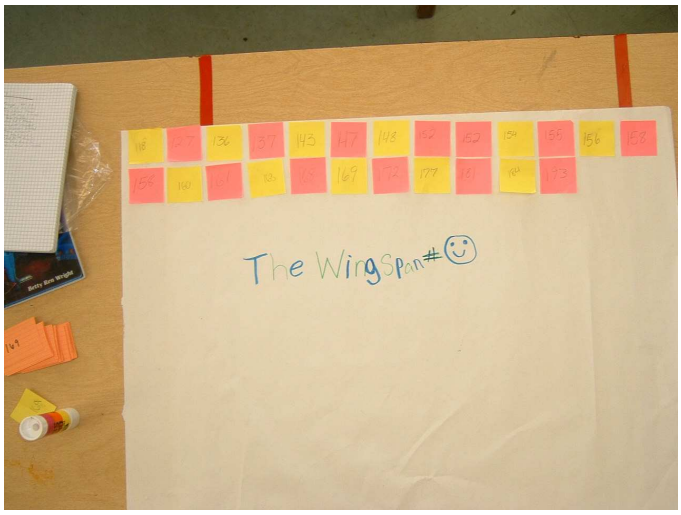

Figure 2. Ordering data as a list.

Others relied on space to convey a visual sense of order. The student solution displayed in Figure 3, a type of array graph (Snecedor & Cochran, 1968), exemplifies the latter. Bars or lines represented magnitudes of measurements. The designers, but not typically other members of the class, indicated that plateaus showed modes or clusters of values.

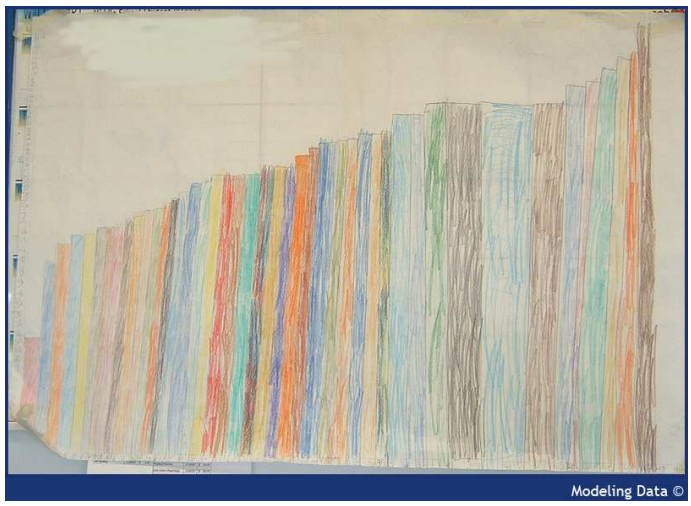

Figure 3. Invented array graph.

Elaborating order. A second class of solutions appeared to elaborate on order by highlighting relative frequency. Figure 4 illustrates this propensity. Students ordered the cases and displayed their relative frequency as a square icon. Note that the interval between case values is not represented. When the teacher asked the students which values would not be likely to recur if they measured again, students pointed to the lowest value. The display made the multi-modal nature of these data visible. The statistics represented on the display are remembrance of past classes – things that one did to batches of data. But after computing them (some incorrectly), they never referred to the statistics again.

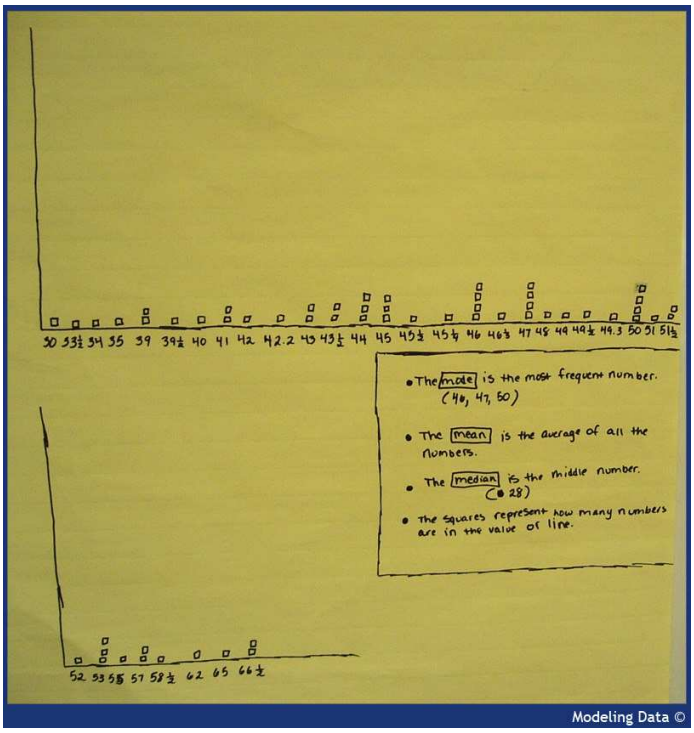

Figure 4. Ordered case frequency display.

Grouping and ordering. Solutions that involved grouping similar values into "bins" or equal-interval groups were relatively infrequent. The designers of the display depicted in Figure 5 grouped measurements in 10s, and they ordered the bins not by magnitude of the measurements but instead by relative frequency. Another pair of designers in the same class rendered their display to coordinate the order of the magnitude of the observed measurements with the relative frequency of each interval class (Figure 6). The corresponding difference in the shape of the data is striking.

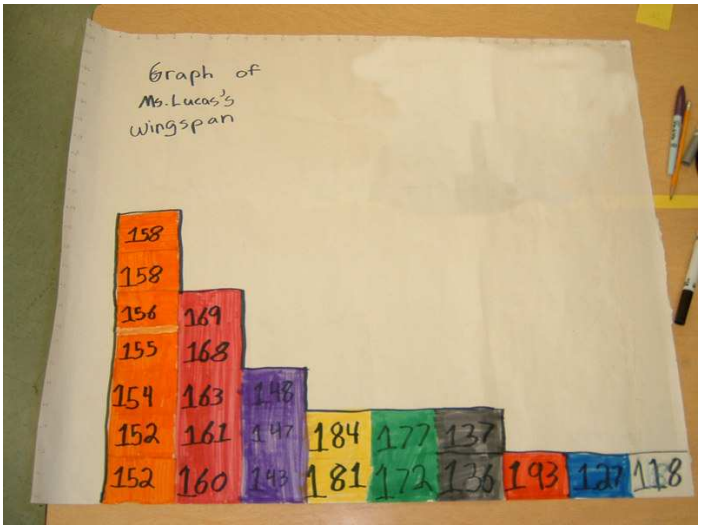

Figure 5. Bin display ordered by relative frequency.

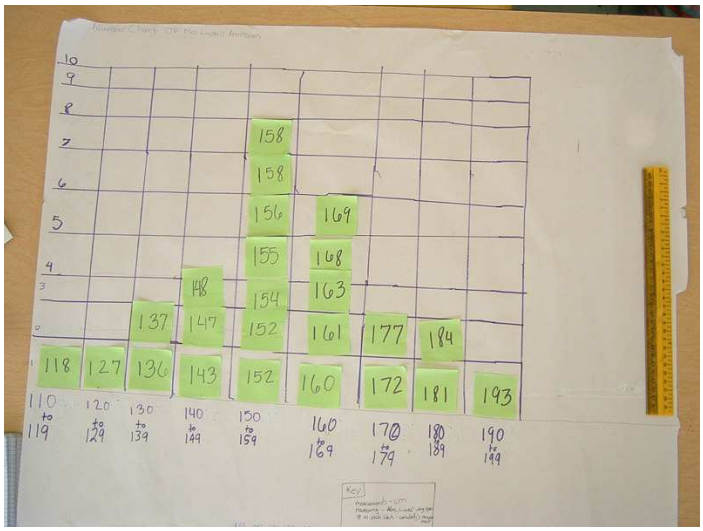

Figure 6. Bin display coordinating case magnitude and frequency.

Interval displays. The least common form of recurrent display was that of interval. These were developed by students who wanted to represent both what was missing as well as what was present in the data, so that holes and clumps could be viewed simultaneously. For example, in Figure 7, a pair of sixth-grade students listed relative frequencies where zero indicated missing values in the interval described by the observed measurements. Hence, 0 = 14 refers to the number of values in the interval between 30 feet and 66 feet for which there was no missing case. The 1 = 9 refers to the number of values in the interval for which there was only 1 case missing. Figure 8, a display designed by a pair of fifth-grade students, illustrates similar attention to interval but in a manner that is more conventional.

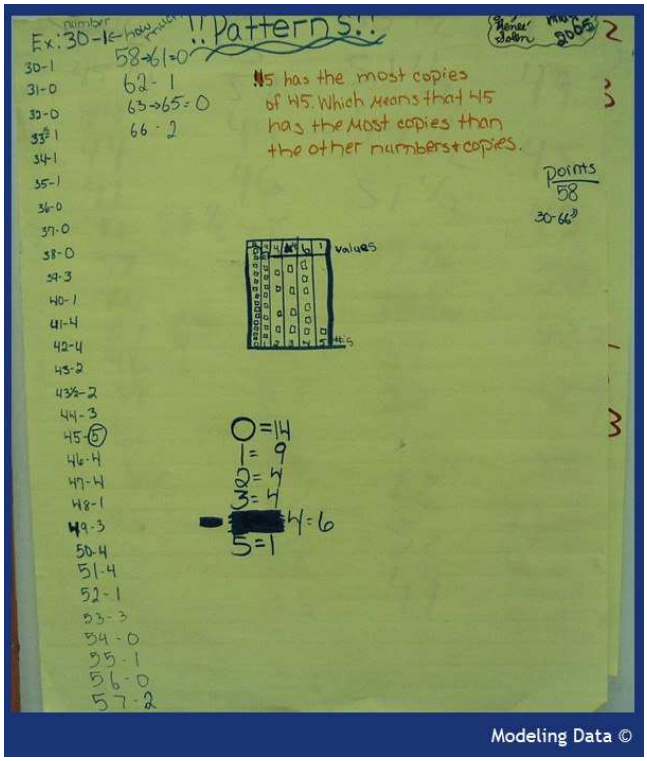

Figure 7. Representing what is missing.

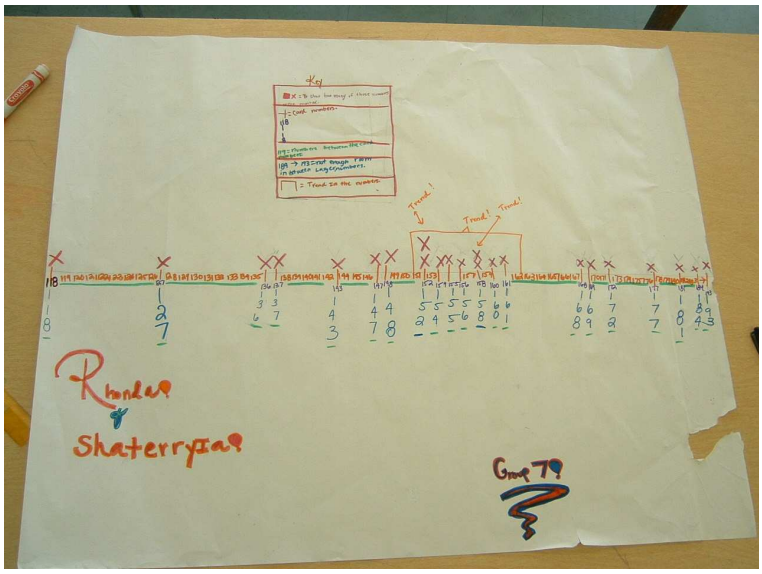

Figure 8. Interval display of relative frequencies.

Comparing representations. Discussions about the variations in design helped students develop an appreciation of different senses of the "shape" of the data. However, students typically focused on individual displays and did not spontaneously engage in comparative analysis. When prompted to compare two different kinds of displays, they often referred to qualities such as icons employed by the designers. For example, students said that they could see squares in one display (to show number of cases) but these were not used in another display. Students often mentioned that a certain display was easy to be seen because it had larger text size. More rarely, a student looked at a display that listed all possible measurements on a number line and said, "They put numbers in between, so you can see how far they went." Hence, I often took a more active role, drawing student attention to trade-offs among displays by asking them to translate a cluster of cases from one representational scheme into another. I also asked students to develop and test conjectures about the relation between the size of a bin (interval) and the resulting shape of the data. These scaffolds appeared to raise students' awareness of relations between design decisions and shape.

Inventing Statistics

Following the invention of a representation of the data, students were challenged to invent a measure of the "best guess" of the length of the attribute (e.g., the height of the school's flag pole). At this point in the learning progression, we anticipated that students could draw on resources of representation and on their knowledge of how the measures were produced. By considering how to develop a measure, we aimed to engage students in deeper consideration of the nature of distribution. What might be worth attending to about the data? Students could use any of the invented displays to help answer this question. We later engaged students in a similar process to develop a measure of the precision of measurements. The definition of precision was intentionally left up to the imagination of the students, so that we could engage students in the relation between measure and qualities of attributes noted in the upper triangular region in the data modeling cycle displayed in

Figure 1. During this period of time, we introduced students to *TinkerPlots* (Konold & Miller, 2005), so that *TinkerPlots* capabilities for dividing and re-organizing the data could be used to construct a measure of precision.

After inventing measures, other students attempted to make use of them. The pedagogical intention was to help students consider the communicative uses of algorithm. Students tried out their methods with other batches of data (to promote generalization), including the measurements of the same attribute with a better tool. For the latter, students noted a reduction in the spread of the data, and I asked if their measure corresponded in meaningful ways to what they could readily perceive in the displays.

Recurrent Patterns of Invented Statistics

Many students struggled with the very idea of inventing of a measure. Some suggested that the only reasonable approach was to ask an authority – a member of the custodial staff or the manufacturer – to find the height of a flagpole. Others found the notion of representing many measurements by a single value implausible. We seized these challenges as opportunities to conduct conversations about qualities of good measures and of the need to be explicit about one's method, so that others could find the same measure.

Measuring centre. Students' invented solutions to estimate the true measure of the attribute generally focused either on repeated values or on the location of the centre clump. Because the data were often multimodal, modal solutions were perceived as less useful, because the inventors typically failed to justify one choice of mode rather than another. Most solutions involving the centre clump used a graphical method to identify the centre clump, and then found the middle value of this centre bin. Many students found this persuasive, but others pointed out that it left out many of the other measurements. A few student teams (at least one in every iteration of the design studies) invented the median, although they did not know this convention at the time of invention. Their reasoning was guided by a sense of splitting the data "in half" and they used bin displays of the data to count an equal number of cases from the tails of the distribution toward the centre. In some data sets, the number of cases was even and the choice for median did not correspond to any observed value. Classmates objected when the median value was not instantiated by an actual measurement, but were persuaded by appeal to the measurement process: The median represented a value that might have easily been someone's actual measurement. It was a "possible measurement". This form of student reasoning signalled a shift away from considering only cases toward considering the aggregate.

Measuring precision. Students' efforts to develop measures of precision most often generated a focus on the "closeness" of the data. More precise measures were those that were closer. We supported this intuition by asking students to predict the value of the measurements if the measures were "absolutely" precise. The three most common solutions to the problem of precision were (a) focus on extreme cases (the range), (b) focus on closeness as distance between a case and other cases or a common point, such as the median, and (c) centre clump solutions, motivated by considerations such as "where the precision was where most people had their numbers".

The range corresponds to convention and thus requires no further explication. The activity of a pair of fifth-grade students exemplifies the second class of solution methods. Their method was spurred by consideration of potentially perfect agreement among the measures, which they suggested would result in no spread or a measure of 0. I asked how

they might define their measure so that zero would result. Their response was to consider differences between each case and the median (which they had invented in the previous portion of this phase of the design study). On the basis of previous work with integers, they decided that they would first find the absolute value of each difference. Then, they proposed finding the sum of these absolute values. Their confidence in this measure was bolstered by its ability to differentiate between distributions of measurements where students employed more precise and less precise tools (e.g., 15-cm rulers vs. metre stick for arm-span). I asked students what they might expect if the number of measurers using the more precise tool increased to 100 (about 3 times the original sample) and this precision was compared to the less precise tool used by fewer measurers. The students noticed that use of their measure would mislead: 'People will think that the more precise tool is worse than the less precise tool' (' denotes paraphrase). To solve this problem, one suggested the modal difference and the other, the median. They settled on the median but had difficulty maintaining the relation between the medians for the distribution of measures and of differences (Figure 9). My suggestion to consider the median of these differences as representing "typical closeness" appeared to stabilize this distinction (meaning that when presenting to classmates, they were able to clearly articulate the distinctions).

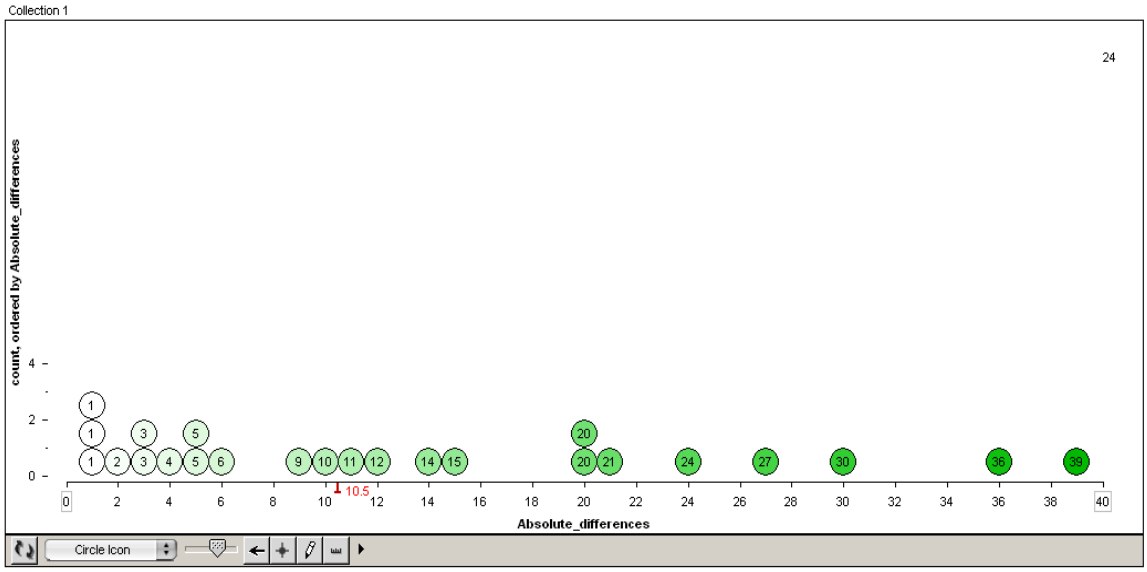

Figure 9. TinkerPlots graph of absolute values of differences with indication of median difference.

Student focus on difference often led to unexpected consequences. For example, one sixth-grader, Robert first focused on the distance between the extreme values of the distribution and the mean. I asked him how he might characterize the precision of the group of measurers rather than just two of them. He decided that he would average the differences, because this would result in a method that would indicate how close the measurements were, "on average". When he attempted to find the mean of the differences, he was surprised that the sum was zero. Robert was puzzled, but he reiterated that he thought his method was good for finding the distances between each score and the mean. He plotted each difference with *TinkerPlots*, and wondered what might have gone wrong (Figure 10).

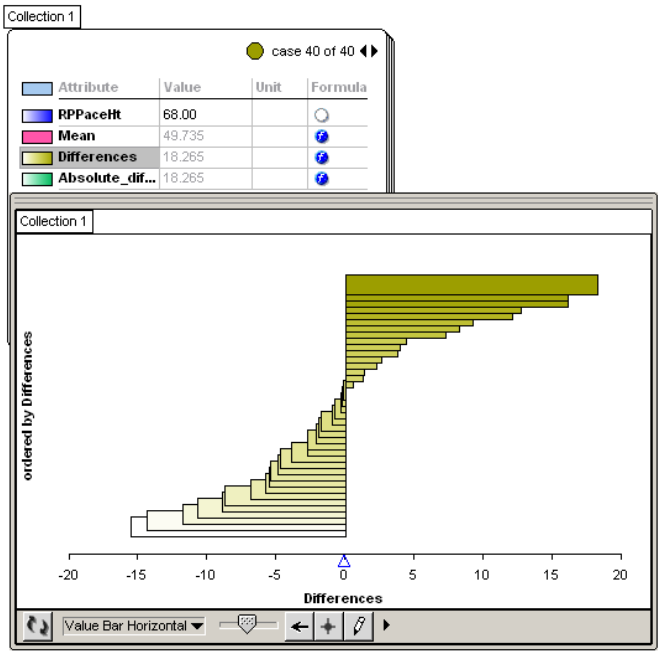

Figure 10. TinkerPlots display of Robert's signed differences.

In light of class discussions about some estimates being over and some under the real height of the flagpole, I asked if Robert were more concerned about the direction, or the magnitude, of each difference. Robert mentioned that the direction of the difference was not that important – some measures *must* be greater than the mean and others less. Hence, what mattered was how far each measure was from the mean. I built on Robert's insight to introduce the absolute value function. Robert used the absolute value function to generate the average deviation. He then plotted the absolute values of the differences, and located their average value – the average deviation (Figure 11), although Robert did not know this convention.

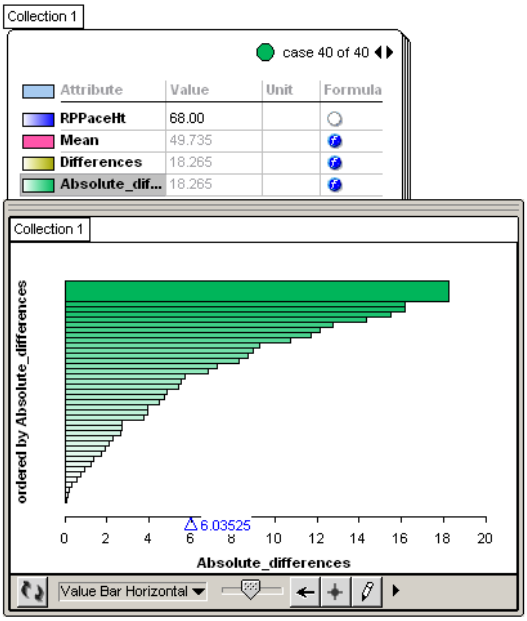

Figure 11. Plot of absolute values of differences and average deviation.

In contrast to close attention to difference, some students defined precision by attending to the relative compaction of the centre clump. Attention to the centre clump typically resulted in measures of precision that corresponded to the inter-quartile range. This definition was supported by the *TinkerPlots* function of "hat plot", but students often used this function only after developing a very similar measure. For example, the solution developed by one sixth-grade student for measuring precision found the lower and upper bounds of the decade-interval that contained the mean. I capitalized on this intuition to introduce the hat plot function, to which the student responded by adding the reference lines to indicate the lower and upper bounds of the mid-50, as displayed in Figure 12.

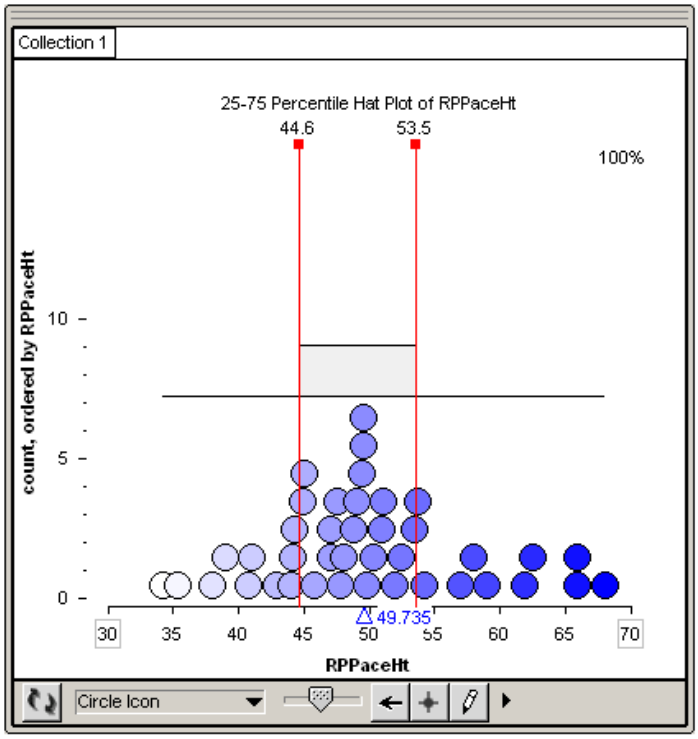

Figure 12. A 25-75 percentile hat plot with reference lines.

Modeling Measure

Following invention of representations and statistics to describe observed trends in variability across different measurement contexts and tools (e.g., arm-span and head circumference, with lower and higher quality tools), the third phase of the learning trajectory is designed to introduce students to the pragmatics and epistemology of modeling chance. We begin with explorations of the conduct of chance devices, starting with hand-held spinners and graduating to a new version of *TinkerPlots* that supports this type of simulation. For example, Figure 13 displays the results of a simulation of a 50-50 spinner with a sample size of 10. Students conducted investigations such as these with varying sample sizes, and we asked students to account for observed differences in departure from expectation as they ran each simulation repeatedly. The line in the Figure 13 was invented by a sixth grade student who thought that changes in slope were a good indicator of departure from expectation as she repeatedly ran the simulation.

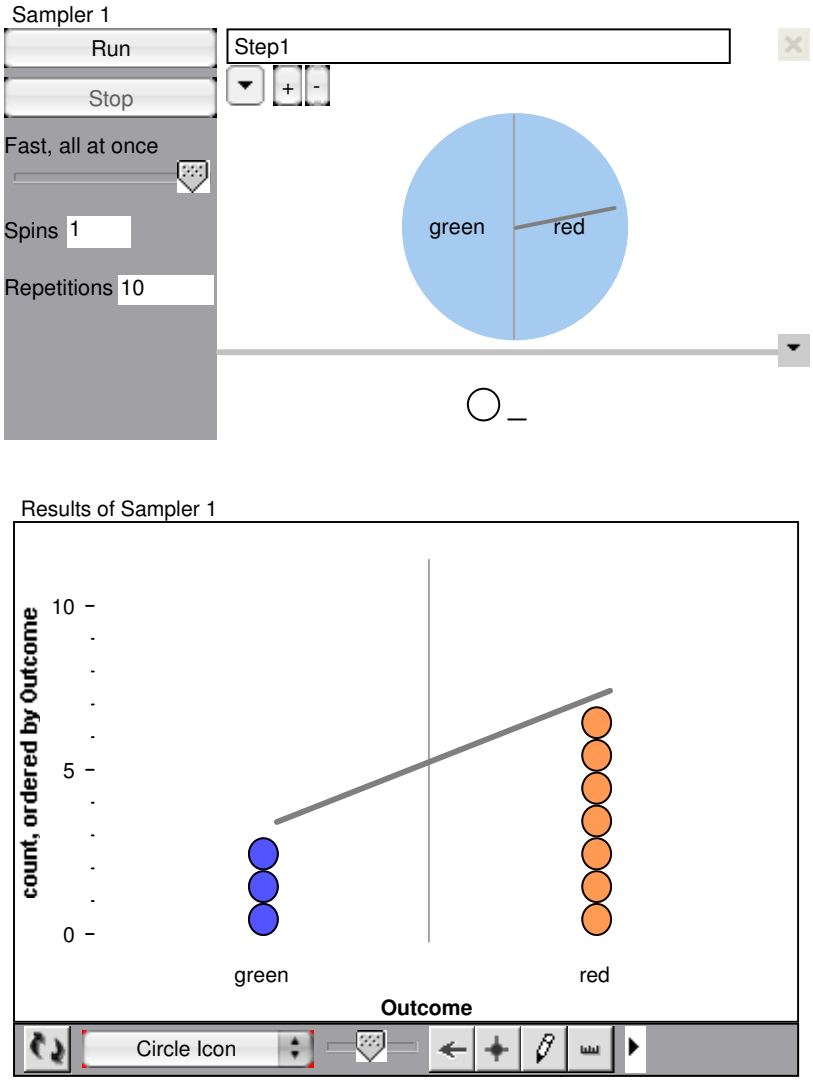

Figure 13. Exploring chance with *TinkerPlots.*

Modelling observed measure. Following investigations of chance, we introduced a prospective model to students of observed measurement as constituted by two sources. Both were familiar to the students. The true measure of the attribute was not directly accessible, but could be reasonably approximated by a centre statistic. The differences among measurements could not be attributed to change in the attribute. (One fifth grader thoughtfully noted that her teacher's head circumference would not change during the interval of measure but she could not say what might happen in the future!) Hence, differences in measure were due to errors of measure. Because students were familiar with processes of measure, we expected that they would be capable of generating conjectures about sources of error. For each source of error, students constructed spinner models that used area to represent relative likelihoods. Relative magnitude and direction of error were also represented as positive and negative values, in the original units of the measure employed by the students. After students constructed and ran simulations of their models, they revised them, as needed. During the final portion of the activity, students constructed "bad" models – models that were designed to employ the same model structure but produce

results that would be judged as poor fits to the observed values. This concluding activity provided a window to students' conceptions of model fit and their skill in using the behavior of chance to create the intended structure of outcomes.

Recurrent Patterns of Modeling

Our approaches and technologies for modeling have been revised during successive iterations, so we are least confident of the stability of results. However, during three iterations of the design studies, students appeared to be capable of readily identifying sources of error. For example, when measuring the arm-span of the teacher, students noticed that use of the 15-cm ruler produced much larger spreads (and less precision) when contrasted to the use of the metre stick. They attributed this difference to needing to iterate with the shorter ruler more often. Each iteration provided an opportunity to produce either over-estimates of the true length or under-estimates. Students attributed the former to "laps", instances where the beginning of one measure and the end of another overlapped, resulting in repeated measure of the same distance. The latter were attributed to "gaps", instances where the end of one iteration and the beginning of another were not aligned, resulting in an unmeasured distance.

To illustrate, I consider the efforts of one pair of fifth-grade students to model the batch of measurements of the circumference of their teacher's head. They designed spinners to correspond to three sources of error, which they termed ruler error, slippage error, and reading error. The first two sources of error referred to potential difficulties using tools to measure the circumference of the teacher's head. For example, slippage referred to the tape slipping or stretching as they wound it around the head. Ruler error referred to the difficulty of establishing a common beginning and ending point for the measurements and for measuring the circumference in exactly the same imagined path around the head. Reading error referred to perceptual difficulties, for example, a measurer might have difficulty judging the number of cm. to the nearest whole number. Each observed measurement was represented by the sum of random error (the sum of the 3 spinners) and the median of the observed measurements, representing an estimate of the true length of the circumference. These spinners are displayed in Figure 14. After running this simulation, the students noticed that it tended to overestimate the centre of the distribution and to produce spreads that were not aligned with the observed values. Hence, they re-designed the spinner depicting ruler error (the far left of Figure 14) to eliminate unrealistically large magnitudes and likelihoods. The resulting simulation was a better match to the shape and centre of the observed values. During the conduct of this simulation, the students noticed that net errors were occasionally zero and that unlikely events nonetheless occurred.

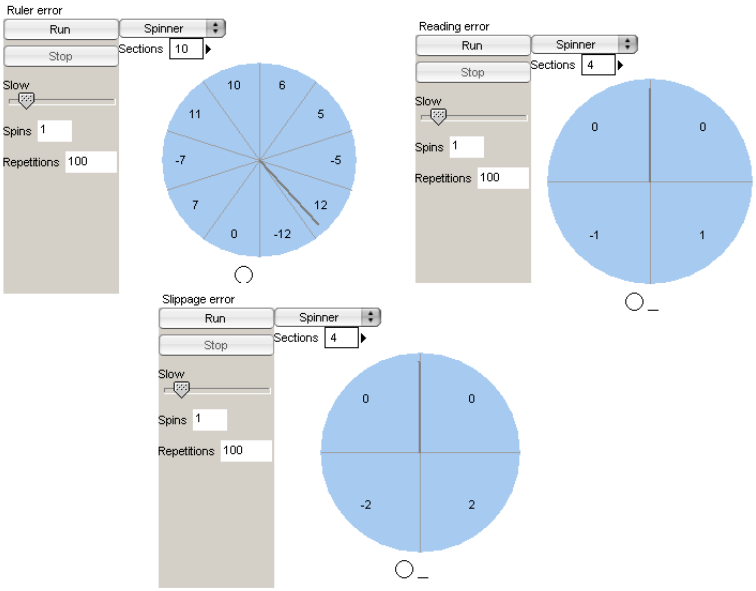

Figure 14. Simulation of sources of random error.

Bad models were a playful way for students to investigate further relations between model design and outcomes. For example, in Figure 15, a fifth-grade student managed to invert the shape of the observed distribution and to produce a skew as well.

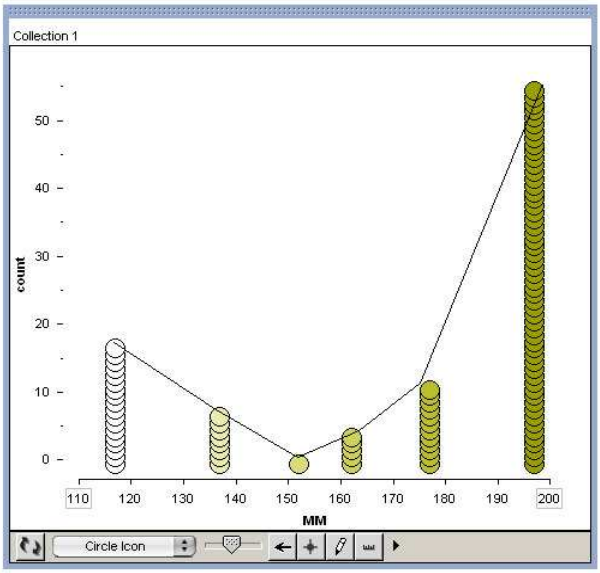

Figure 15. Results of a simulated bad model of normally distributed measures.

Designing Assessment to Support Instructional Design

In most design studies, day-to-day decisions are made in light of evidence about student thinking, most often obtained from inferences based on students' discourse and artifacts that they produce. Much of the previous presentation of recurrent patterns of student reasoning follows in this tradition. In design research, assessment is often considered after the fact, as summative evidence of more widespread patterns of individual performance. However, in our design studies assessment played a central role, both in the conduct of the studies and in the interpretation of the results. In fact, one of the anticipated outcomes is the creation of an assessment system.

To create an assessment model, our conjectures about the forms of knowledge and the nature of conceptual change underpinning learning about variability were expressed as *progress variables* (Wilson, 2005). Progress variables model trajectories of development. They demand that designers of learning progressions make their commitments about conceptual growth explicit. We constructed progress variables in seven conceptual strands: (a) theory of measure (conceptual landmarks for understanding the nature of units and scales of measurement, which are prerequisite understandings for the learning progression), (b) modeling measurement, (c) data display, (d) meta-representational competence, (e) concepts of statistics, (f) probability/chance and (g) informal inference. Table 1 illustrates a summary of the Data Display progress map, which lays out our conjectures about prospective transitions in students' conceptions, from case-based to aggregate-based ways of constructing and interpreting data displays. The full version of each construct contains examples of each performance in both text and video formats.

Although progress maps may appear to have a preordained character, in fact, they are negotiated as the design study unfolds, so that progress maps take several design iterations to "settle". Hence, they serve as a visible trace of prospective conceptual landmarks for the design team.

Based on the construct maps, we designed items to support instruction and to index student progress over longer periods of time. To support instruction, some items were designed as formative tools to diagnose student conceptions. These were administered as weekly quizzes, and the results were employed to re-design the learning progression. For example, during one design study, the results of a formative assessment indicated that many students interpreted their classmates' invented statistic, the median, to be a half-spit of the data located in the "middle" of a string of data. They apparently did not consider the order of the data as critical, relying instead on the spatial centre of the data presented by the inventors of the statistic. Consequently, we decided to problematize "half" by contrasting the distance-based image of the mid-range with the count-based definition of the median. Students thought that any estimate of the best guess of the length of the arm-span should be located in the centre clump. Their image for mid-range was a paper strip folded into two congruent lengths, an image familiar to them from class work earlier in the year finding partial-units of length measure. The fold line of this strip located ½; but, what was the relation of this distance-based sense of half to the half demarked by the median? If the mid-range was "halfway", how could the median also be considered half? How could counting result in a location in the centre clump? We constructed several small sets of imagined measurements with the lowest or highest values located in the centre. By simply counting, the extreme values were considered best guesses of the true measure. Yet, this contradicted children's sense. This contradiction was resolved by re-examining the role of order in

determining the median, and by juxtaposing two different senses of "1/2-split" – one based in distance and the other in position within an ordered sequence. We also took this opportunity to investigate robustness of the statistics proposed – by investigating the effects of "one bad measurer" on the estimate of true measure. (The mid-range declined in popularity when students considered that just one student-measurer could shift the value of the mid-range out of the centre clump.) These modifications were incorporated into subsequent iterations of the design.

Table 1
The Data Display Construct

Level	Performances
DaD6. Integrate case with aggregate perspectives	DaD6(a) Discuss how well individual values or regions represent the patterns seen in the whole distribution, or vice versa.
DaD5. Consider the data in aggregate when interpreting or creating displays	DaD5(b) Quantify aggregate property of the display using one or more: ratio or proportion or percent. DaD5(a) Recognize that a display provides information about the data as a whole that goes beyond any of the cases by themselves.
DaD4. Recognize or apply scale properties to the data	DaD4(b) Recognize the effects of changing bin size on the shape of the distribution DaD4(a) Display data in ways that use its continuous scale (when appropriate) to see holes and clumps in the data.
DaD3. Create categories of cases based on relationship among them	DaD3(c) Identify data points that are dissimilar to the rest. DaD3(b) Identify grouping of similar values (e.g., high, medium, low values). DaD3(a) Note similar values or "clumps" in the data set.
DaD2. Concentrate on cases when working with data	DaD2(b) Manipulate data attending only to its ordinal properties. DaD2(a) Concentrate on specific data points (minimum, maximum, median, mode), without relating these to any structure in the data.
DaD1. Treat data as collection of individual numbers or attributes	DaD1(a) Manipulate, notice and explore qualities or relations of data values, without relating to the goals of the inquiry.

Although this effort is still a work in progress, we are currently working to articulate an assessment system that will span both instruction and accountability. From the perspective of conducting studies of learning, the formative assessments standardize our commitments about what counts as evidence of student reasoning. The summative assessments provide a less fine grained but broader spectrum to track conceptual change. This provides an opportunity to engage in design experiments, in which the implications for learning of different instructional designs can be contrasted in a common metric.

Discussion

The links between data analysis, chance, and modeling have often been severed in school mathematics. Yet, in a wide variety of professions, data modeling is integral to practice. The epistemology in professions is one of model building and competition, not

one of "descriptive" statistics, followed by "inferential" statistics, which is the standard practice in schools. I propose restoration of the link between data modeling and statistical reasoning in schooling, not merely because it is what professionals do, but more importantly, because it is a viable and fruitful approach for supporting the growth and development of student reasoning about variability. Variability is ubiquitous and it is critical for thinking in the 21st century that we equip students with ways to reason about it.

The learning progression outlined in this paper rests on several general principles of learning and on the potential affordances of measurement as a context for investigating variability. The first is that of agency. If measure is framed as activity, rather than as a product, students can mentally simulate the role of agents and/or they can literally enact measurement process. Agency mediates student apprehension of variability by making process transparent (e.g., individual measurers can recall qualities of method and measure that might lead to "mistakes" in measurement), and it grounds symbolic expression, in that students can readily relate presentational qualities (e.g., hills in graphs) and measures thereof (e.g., medians as measures of centre) to specific forms of activity. A related virtue of agency is that qualities of distribution can be viewed as emerging from the collective activity of agent-measurers. Hence, a statistic, such as the median or mean, can be viewed readily as a measure of central tendency (Konold & Pollatsek, 2002), and the explanation for such a tendency can be attributed to the notion of a true or fixed value.

Second, developing representational and meta-representational competencies have important conceptual consequences. The diversity of representations invented by students supports the concept that the shape of the distribution is not a Platonic ideal, but rather, a result of a particular set of choices made about what to attend to, and what to obliterate, in a system of representation. Not all students fully grasp the idea of representational trade-off, but supporting comparisons among representations provokes mathematically fruitful consideration of different meanings of the "shape" of the data. Seeing hills and valleys is one thing, knowing how they are produced and how they might be magnified or even eliminated is another. We strive for the latter, and it appears that this is a consistent outcome when we deliberately instigate comparisons among representations.

Third, inventing measures of what students can readily "see" in a set of data invites closer inspection of the qualities of the data that contribute to the perception. Students' invention of measures of centre and spread support consideration of just what one might mean by each. Thus, there is an intimate relation between conceiving of the "centredness" or "spreadness" of the distribution and its measure. What students see after inventing measures is often different than what they saw before such invention. Thus, measure is an important cornerstone to quantification (Lehrer, Carpenter, Schauble, & Putz, 2000; Thompson, 1994). Inventing measures supports a meta-conceptual development: What does it mean to measure and what are qualities of good measurements? These developments are supported when students employ their inventions to measure the attributes of new distributions that were formed when measurers used different methods or tools. For example, students' experience suggests that measuring the arm-span of a person with a 15-cm ruler is more error prone than the same measure employing a metre stick (fewer iterations lead to less error). Hence, it makes sense that the distributions have different precisions and that the measure ought to reflect these differences. Measure allows too for a new form of inquiry not as readily sustained by the eyes: How much more (or less)?

Fourth, the conceptual landscape of modeling is altered by the technologies deployed for modeling. When we first began, students used hand-held spinners to construct models, and these were certainly adequate tools for engaging in the process of modeling chance. However, we cannot help but notice that the introduction of *TinkerPlots* alters this landscape. One form of alteration is in ease of model design and revision. Although we wish for more capability from *TinkerPlots*, and we are confident that we will soon see it, the current implementation allows for much more rapid prototyping and running of models. We believe that this has a conceptual consequence: Models that are run more often invite attention to sample-to-sample variation in outcomes. This embarks students on the road to sampling distribution, an unintended consequence from the point of view of our initial conception of the learning progression.

Last, although we often hear that cognitively guided assessment is a virtue, it is difficult to find many examples. Of course, virtue is always distributed more like Poisson than Gauss, but our work with colleagues at the Berkeley Evaluation and Assessment Research Center and the work of Jane Watson and her colleagues (e.g., Watson, Callingham, & Kelly, 2007) suggest that linking assessment to models of learning statistics is not a trivial pursuit. When we work collaboratively with assessment experts during the design of instruction, we find that both of our professional worlds are enriched, and we hope, so too are those of the students.

I conclude with a lamentation. The opportunities for supporting student reasoning about variability are often confined to mathematics education. Yet the origins of the mathematics of variability arose in contexts of modeling nature, and these contexts are still a primary arena for modeling variability. Unfortunately, school science works full time to hide this variability from students, especially in pursuit of laboratory exercises with gargantuan effect sizes that render inference moot. This is a lost opportunity. A science education that encouraged student inquiry and model development would be a natural site for grappling with issues of variability.

Acknowledgement. The research described in this paper is collaborative. Thanks to Min-joung Kim, Charles Munter, Leona Schauble, and Wenyan Zhou for their valuable contributions to the design and implementation of instruction. Mark Wilson and Tzur Karletiz of the University of California-Berkeley lead our effort to design the assessment system. Cliff Konold of the University of Massachusetts-Amherst supports student modeling of chance with his re-design(s) of TinkerPlots.

References

diSessa, A. (2004). Metarepresentation: Native competence and targets for instruction. *Cognition and Instruction*, *22*(3), 293-331.

Giere, R. N. (1992). *Cognitive models of science.* Minneapolis, MN: University of Minnesota Press.

Greeno, J. G., & Hall, R. (1997). Practicing representation: Learning with and about representational forms. *Phi Delta Kappan, 78*(5)*,* 361-367.

Hall, R., Wright, K., & Wieckert, K. (2007). Interactive and historical processes of distributing statistical concepts through work organization. *Mind, Culture, and Activity, 14*(1&2), 103-127.

Konold, C., & Miller, C.D. (2005). *TinkerPlots: Dynamic data exploration*. [Computer software] 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.

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

Lehrer, R., & Schauble, L. (in press). Contrasting emerging conceptions of distribution in contexts of error and natural variation. In M. Lovett & P. Shah (Eds.), *Carnegie Symposium on Cognition: Thinking with data.*

Lehrer, R., Carpenter, S., Schauble, L., & Putz, A. (2000). Designing classrooms that support inquiry. In J. Minstrell & E. H. van Zee (Eds.), *Inquiring into inquiry: Learning and teaching in science* (pp. 80-99). Washington, DC: American Association for the Advancement of Science.

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

Snedecor, G. W., & Cochran, W. G. (1967). *Statistical methods*. Ames, IA: University of Iowa Press.

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

Thompson, P. W. (1994). The development of the concept of speed and its relationship to concepts of rate. In G. Harel & J. Confrey (Eds.), *The development of multiplicative reasoning in the learning of mathematics* (pp. 179-234). Albany, NY: SUNY Press.

Watson, J. M., Callingham, R. A., & Kelly, B. A. (2007). Students' appreciation of expectation and variation as a foundation for statistical reasoning. *Mathematical Thinking and Learning, 9*(2), *83-130.*

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry (with discussion). *International Statistical Review*, *67*(3), 223-265.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.