Constructing Data, Modeling Chance in the Middle School

Richard Lehrer[1], Cliff Konold[2], & Min-Joung Kim[1]

Vanderbilt University[1], University of Massachusetts-Amherst[2]

Paper presented at the 2006 annual meeting of the American Educational Research Association, San Francisco, CA.

Abstract

We 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. Both of these developments have historic parallels, and the intricate relation of measure and data are also key components of ongoing professional practice (see Hall et al., this symposium). From a learning perspective, the learning trajectory reflects a commitment to several related principles: (a) constituting a learning progression as encounters with a series of *problematics*; (b) *representational fluency* and *meta-representational competence* as constitutents of conceptual development in a discipline; (c) *invented measures* as grounding students' understanding of statistics and (d) an *agentive perspective* for orienting student activity, according to which distribution of measures emerges as a result of the collective activity of measurer-agents. Instructional design and assessment design (see Wilson et al., this symposium) 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.

The discipline of statistics had its origins in problems of modeling variability (Porter, 1986; Stigler, 1986). History hasn't changed all that much: Professional practices of statisticians invariably involve modeling, and as in other sciences (see Giere, 1988), it is through model contest that statistical concepts become more widespread and stable (Hall, Wieckert, & Wright, this session). Another lesson of history is of particular importance to us: 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 errors varied. Mathematical efforts to characterize the form and structure of measurement variability gave rise to concepts and models still in use today, such as "least squares fit." We aim to exploit these historic relations by introducing children to modeling variability in contexts of repeated measure.

Affordances of Measurement for Mathematizing Variability

Our choice of measurement as an entrée to variability is not grounded in mere homage to history but in rich prospects for learning. One prospect is that measurement affords *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. As a consequence, foundations of statistical inference, such as notions of repeated, random process (Liu & Thompson, 2002), have counterparts in distinguishable forms of activity—forms that students can readily identify. For example, stochastic process relies on the (mental) construction of trial (Horvath & Lehrer, 1998)—the assumption of identity over repeated instances of a process. In measurement, trial finds expression as the repeated activity of a measurer, or in the collective activity of a group of measurers acting in concert. Qualities of measurement, such as its precision, have observable

consequences for variability. For example, students can readily notice the difference in variability when they use more or less well-machined tools (Petrosino, Lehrer, & Schauble, 2003). 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 its symbolic expression, in that students can readily relate presentational qualities (e.g. hills in graphs) and measures thereof (e.g., medians as measures of center) to specific forms of activity.

A related affordance of measurement is emergence. Presentational and representational 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. Similarly, a statistic summarizing the spread of the distribution can be readily interpreted as reflecting tools or methods employed by agent-measures. A shape, such as the "normal" curve, can be thought of as the collective result of measurers who over-shoot and under-shoot the true measure, but usually not by very much. Thus, measurement appears to offer promising connections between case and aggregate views of the data. We recognize that much has been made of the prospective problems that emergence poses for theory development. Resnick (1997) argues that it is difficult to understand emergence because of a bias toward expecting that structure must be determined by a central controller, and Chi (2005) suggests related obstacles, chief among which is a difficulty reconciling the ontology of different levels of an emergent process. Chi suggests that emergent processes are (mis)conceived as "direct" processes. However, we consider emergence as a resource for, rather

than as a menace to meaning, because we believe that students who understand how agents function are less likely to confuse different levels or assume that there must be an agent-in-charge.

Third, although each agent is individual, given similar means and tools, agents tend to produce similar measures of the same attribute. "Close, but not identical" is a potential entrée to the idea of interval, creating the potential to view measures within the interval as exchangeable. Interval is foundational to the notion of density, which transforms data into ordered counts within (ever decreasing) intervals. Distributions describe the density of data, and a wide range of studies suggests the importance of distribution to statistical reasoning (Cobb, McClain, & Gravemeijer, 2003; Petrosino et al., 2003; Saldanha & Thompson, in press).

Fourth, measurement affords an entrée to sampling: What might happen if the measurement process were repeated? Because measurement processes are more transparent to students than those in other contexts, such as the genetic recombination responsible for much of "natural" variation, we believe that students are likely to understand that some regions of the measured values are more likely to be reproduced than others. For example, measurer-agents tend to come close to the true value of the measure, reproducing central tendency from sample-to-sample.

Designing for Learning to Mathematize Variability

We designed a sequence of tasks and tools that exploited these contextual affordances with the aim of supporting students' efforts to develop a mathematical system for describing variability. Our design was guided by heuristics that shaped but could not determine our choices.

Problematics. First, we conducted what Thompson (2000) refers to as a conceptual

analysis. In this case, we made conjectures about concepts and practices that might play fruitful roles in developing the mathematics of variability. As we describe later more fully, measurement appeared to offer prospects for children to interpret variability by recourse to processes that would be meaningful to them (Konold & Pollatsek, 2002; Petrosino et al., 2003). We imagined situations and tasks involving measurement in which foundational concepts, such as distribution and statistics, would be experienced as *problematic*. For example, many studies of statistical reasoning examine students' conceptions of the mean. It was our intention to create a context in which students would grapple with the rationale and the need for a measure of center, rather than to prescribe the mean as a solution to a problem that they did not yet experience as meaningful (see Lesh, Hole, Hoover, Kelly & Post, 2000, discussion of principles for design of model development sequences). The outcome of this analysis was a series of problematics, each linked to a particular form of activity. For example, as we describe next, we were concerned that students not treat shape as an ineluctable quality of the data, so we developed activities in which different senses of shape might conceivably arise.

Meta-representational competence. Students invented ways to inscribe/represent data so that others could notice what they viewed as important qualities of the data. This tactic was intended to foster representational fluency (Greeno & Hall, 1998). We anticipated that by inventing displays of data, students could recognize that visible qualities of data, such as its shape (which is often treated transparently in traditional curricula), are a consequence of representational choices, and not only qualities of the data. We also asked students to compare and contrast their invented displays, with an eye toward promoting meta-representational competence (diSessa, 2004). We anticipated that comparing and contrasting student inventions

would clarify relations between visibility of qualities of data and representational choices. Some displays highlighted some features of the data, even as they obscured others. We were especially interested in helping students understand how displays that grouped data worked to produce a shape characteristic of many measurement situations (the "normal"), because this shape could be readily related to measurement processes and measurement errors.

Invented measures. Third, students invented measures of the qualities of distribution as a means to render statistics as sensible summaries. When inventing measures, students consider which aspects of the distribution are worthy of attention (Petrosino et al., 2003; Schwartz & Martin, 2004). Such attention focuses students on the role and function of statistics; armed with such knowledge, they can come to appreciate the problem and the trade-offs that conventions, such as the mean or variance, represent.

*Changing the representation*al *landscape*. Students first used familiar paper-and-pencil tools. Our intention was to rely on these traditional means to make certain aspects of data, such as the shape of the data, problematic. That is, we did not want to hand students solutions to problems they had not yet experienced. After students had the opportunity to invent representations and critique those developed by classmates, we introduced TinkerPlots™. TinkerPlots™ alters the representational landscape by introducing new notational systems that are difficult, if not impossible, to create with paper-and-pencil. For example, TinkerPlots™ dynamically links intensity of color to quantity, making trends in the data easier to spot.

Framing the Learning Progression

With these heuristics in mind, we framed a prospective sequence of tasks and tools that could serve to introduce students to modeling measurement variability. First, all students

measured the same object with the same tool and method. We anticipated that students would expect all measures to be identical, thus grounding the investigations to follow in (mild) surprise. Why didn't everyone get the same measurements?

Students then designed a visual display of their measurements. Our intention was to instigate exploration of potentially diverse senses of the "shape" of the data. We engaged students in a design critique. Students analyzed how their design choices highlighted some features of data while placing other aspects in recession, and we promoted explicit comparisons between designs. As we mentioned previously, design critiques were intended to support the development of meta-representational competence--helping students come to see different types of display as embodying trade-offs. We privileged shapes of data that resulted from considering counts of data (i.e., the number of cases within a specified interval). As we suggested earlier, our intention was to eventually support students' reading of a distribution of values as a density.

Following the design critique, students invented measures of the "best guess of the real length" of the attribute (e.g., arm-span or height of a flagpole) and of the precision of the data (e.g., tendency for values to agree). By comparing distributions of measurements obtained with less- and more-precise tools, students could explore the behavior of their invented statistics. For example, when the variability decreased (more precise tools), did the invented statistic also decrease? Our intention was to support a view of statistics as descriptors of qualities of distribution (i.e., center and spread) and hence as ways of describing aggregates of data, but not single values. Moreover, the context afforded links between process and distribution—as tools of measurement changed, the center remained stable but the variability changed.

In the next phase of the instruction, students modeled their intuitions about sources and magnitudes of error by employing chance devices—spinners. Observed values were modeled by

summing the "true" measurement and chance errors of measure. Our intention was to introduce the very idea of modeling as a way of making explicit students' intuitions about chance "mistakes" in measurement. Along the way, we anticipated that students would also have the opportunity to consider relations between the outcomes of aggregates of trials and single trials.

The instructional sequence concluded with putting knowledge about distribution to use in making inference. Students measured the apogees of model rockets in flight, created a distribution of these measures of apogee, and then compared this distribution to another distribution of measures taken with a modified rocket (the rocket's nosecone was pointed instead of round). Students considered whether or not the modification made a difference in light of measurement variation. This concluding phase of the instruction introduced students informally to statistical inference.

Investigating the Viability of the Learning Progression

We investigated the viability of the proposed learning progression by examining children's development of the mathematics of variability. Our investigation focused on changes in the forms of and resources for reasoning, and also the conditions under which development can best be supported. Thus, the investigation had the general form of a design study (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003).

Design and development of an assessment system. 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 gestures. Assessment is often considered after the fact, as summative evidence of more widespread patterns of individual performance. However, in the design studies we report, 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. Our colleagues from the BEAR Center at the University of California, Berkeley, will describe interactions between research and assessment more fully, so our purpose is to describe briefly how assessment aided our investigation of the viability of the proposed learning progression.

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*. Progress variables model trajectories of development. They demand that designers of learning progressions make their commitments about conceptual growth explicit. To date, we have constructed progress variables in 7 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 displays, (d) meta-representational competence, (e) concepts of statistics, (f) probability/chance and (g) informal inference. Figure 1 illustrates 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.

Figure 1

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.

The other components of the assessment system also support design iteration. Design studies feature adjustments to instruction based on evidence, and *formative* assessments

standardize the nature what counts as evidence. As our colleagues will describe more fully, the formative assessments in this system summarize prospective instructional contingencies by suggesting next instructional steps based on inference about current states of knowledge and performance. Finally, summative assessments fulfill their traditional role in design studies. They encapsulate the kinds of understandings that we hope students will develop, and they provide a means for examining how widespread these understandings are in a sample.

Design Studies

To date, we have conducted three design studies in a Nashville public school with a predominantly minority population (76% African American, Hispanic, and Kurdish). The school serves about 300 students in Grades 5-8; 69% qualify for free or reduced-price lunch.

The initial design study featured 6 students selected by their teachers in grades 5 and 6 to participate in a daily afternoon enrichment period. Criteria for selection varied by teacher, but the intent was to represent a wide range of mathematics achievement. All students were volunteers whom their teachers judged would benefit by participating. Students met three times a week for approximately three months. Each lesson was 45 minutes. The second design study was conducted to verify trends observed during the designing display phase of the first study. It was conducted in one 6th grade classroom for ten days. The third design study is ongoing in one fifth-grade classroom. To date, all lessons have been taught by a researcher (RL), with occasional assistance from the classroom teacher. Sources of data include digital video, student artifacts, and clinical interviews, collected during these multiple iterations of design.

Our purpose today focuses on (a) students' inventions of representations and their implication for coming to understand varying senses of the "shape" of the data; (b) students' explanations for the effects of changes in measurement processes (which are experienced by

individuals) on distribution (which is a collective and emergent outcome); (c) students' invention and revision of measures of central tendency and variability; and (d) students' models of random errors of measurement.

Inventing Representations

After each student measured the height of the school's flagpole or the length of one person's arm-span, the measures obtained were collected and posted on a whiteboard in front of the room. Students worked in small groups to design displays that showed, without words, all of the data and any of the trends about the data that they noticed. Students posted their displays and classmates, but not the designer, attempted to interpret their meaning. What was the display trying to show? The teacher employed a language for fostering meta-representational competence by asking the class to consider which features of the data were highlighted by a particular display and which it made less obvious, or even hidden from view.

The most striking aspect of this phase of the instructional sequence was the variability of student displays. We present a sample of these displays to illuminate a few themes that we observed across the three design studies.

Highlighting order. Many students structured data by ordering. Some solutions were lists (e.g., Figure 2), but 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 2, 3

Figure 4 is a hybrid of list and space. It orders the data and makes use of higher = greater value. The designers mentioned that they wanted to evoke an image of climbing stairs.

Figure 4

Elaborating order. A second class of solutions appeared to elaborate on order by highlighting relative frequency. Figure 5 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 5

Constructing interval. Some students decided that what was most interesting about the data were the measurements not there: the possible, but not realized measurements. To represent this space of possibilities, they constructed displays that highlighted the missing values, displayed in Figure 6. The frequencies listed in Figure 6 included "0," 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 case. The 1 = 9 refers to the number of values in the interval for which there was only 1 case.

Figure 6

Other student-designers employed a continuous axis to make the gaps or holes in the data more visible. One exemplar is displayed in Figure 7. Here fifth-grade students drew a number line from 118 to 193 and used size of texts and colors to differentiate acquired measurements and missing values. Also, they underlined missing values with green lines, to highlight holes in the data.

Figure 7

Grouping and ordering. In the first two design studies, solutions that involved grouping similar values into "bins" or equal-interval groups were relatively infrequent, and nearly always involved some form of teacher assistance. However, in the most recent design study, fifth-graders who interpreted and constructed frequency graphs in prior social studies lessons predominantly developed graphs with groups of values. However, these graphs often departed from conventions. Designers of Figure 8 grouped measurements in 10s, but they ordered bins based on heights of the bins to highlight which bin had the most measurements. Another group of students divided all measurements into three groups. As a result, bin sizes were not consistent. However, they highlighted that half of measurements were in the middle bin.

Figure 8

Idiosyncratic. Idiosyncratic solutions were created just once and did not seem to have counterparts in other student solutions. For example, one team of fifth-grade students initially ordered the data and then separated it into two groups, based on the parity of the measurement (even, odd). When questioned, they could not relate their choice of data group to any goal. They seemed to treat the data as numbers to be investigated, rather than as data.

Two sixth-graders went to great pains to declare that there was no pattern to the measurements (see Figure 9). The students declared that the numbers did not "help each other." They meant that the numbers could not be described by any rule that they could detect (They gave examples which included the definition of even vs. odd). They also objected to the possibility that data could exhibit pattern, because their previous experiences of pattern featured pattern blocks "square, triangle, circle." The absence of a determinate rule appeared to imply an ontological category of "not mathematics."

Figure 9

Comparing Representations

Discussions about the variations in design helped develop an appreciation of different senses of the "shape" of the data. The activity structure was one wherein students commented about what a display showed about the data and what aspects of the data remained less visible or

even hidden from view. The aim was to foster a sense of trade-off rather than good vs. bad displays.

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."

Yet the diversity of display offered many opportunities for the teacher to highlight other fruitful comparisons (from a mathematical point-of-view). For example, the construction and use of *interval* was made more visible, simply because some student solutions accounted for gaps and holes in the data while others ignored them. Some student solutions included groups or "bins" of values, which resulted in a very different sense of the shape of the data when contrasted to array graphs or lists. However, they might also juxtapose them without regard to the entire range of the interval. That is, students arranged values in order, such as 10's, 20's, and then juxtaposed 40's, when there were no values in the 30's. The resulting display highlighted clumps of values but made "holes" in the data invisible.

Inventing Measures: Center

Students invented measures of the "best guess" of the actual height of the school's flagpole or the real length of the arm-span of their teacher. In this measurement context, this estimate is the center of the distribution.

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. Students' solutions generally fell into three classes: (a) convention, fueled by prior knowledge but not well related to qualities of the data; (b) reasoning about repeated values; and (c) reasoning about the center-clump.

Convention. In the first two design studies, four students calculated the mean because that was what one did with a batch of data: "find the average." They were unaware of any of the properties of the mean, so we altered instruction to make some of its qualities more visible. For example, students investigated the effects of extreme values on the mean. These students did not entertain other conventions, such as modes or medians. In the current design study, no fifth-grade student attempted to find the mean.

Repeated values. Many students reasoned that if two or more people agreed about a measured value, then that value was more likely the "right" one, even when the data were multimodal. For example, several groups of fifth-grade students (in the third, current design study) treated the problem of multi-modal data by asking prospective users to chose the "reasonable" one. Yet, they could not communicate any criteria for establishing reason.

Center clump. The majority of students who did not have a prior orientation to the mean invented analogs to the median. Their reasoning was guided by the appearance of a center clump in the data, when the data were grouped and ordered. Students were attracted to the relative frequencies of the values in the data, likely because these values literally occupied more space

(e.g., they were higher). Most solutions involved finding the middle value of this center bin, a workable solution for measurement data. However, a few groups independently invented the median, guided by a sense of middle as splitting the data into two parts of equal count. 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 signaled a shift away from considering only cases toward considering the aggregate.

Infrequently observed solutions/Hybrid solutions. One pair of fifth-grade students invented a method that placed data into bins, determined the percentage of cases within each bin, deleted the measures in the low percentage bins, and used the mode of the remaining cases as the best estimate of the true measure of arm-span. Another fifth-grade student suggested the midpoint of the distance between the lowest and highest values, the mid-range, as the estimate of the true length of the attribute.

Juxtaposing different senses of half. The results of an ongoing formative assessment (a few item quiz) suggested that many students interpreted their classmates' solution of 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 center of the data presented. 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 center 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 part-units of length measure. The fold line of

this strip located _. 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 center clump? We constructed several small sets of imagined measurements with the lowest or highest value, which students readily suggested were mistakes, in the center of the listed values. 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 center clump.)

Inventing Measures: Precision

Students invented measures of the "precision" of their measurements. Precision was intentionally not well defined, so that students would be put into a position of attempting to describe it. One sense of precision that many students developed was "closeness." Relative closeness of the data could be defined in many ways, and students were encouraged to articulate just how they might measure closeness. TinkerPlots™ was available to students during this phase of the instruction. In the first and third design studies (the second did not include this phase of instruction), students' invented measures were anchored either in conceptions of distance, or in conceptions of the relative compactness of the center clump.

Distance-based solutions. The range was the most common invention and was constructed primarily by the youngest students (the fifth-graders). In an attempt to address all of the observed measurements, one fifth-grader proposed to order all the data, find the successive pair-wise differences, and then find their sum as a representation of precision.

A sixth-grade student (Robert) first focused on the distance between the extreme values and the middle of the distribution. He used the mean to represent the middle. His teacher upped the ante of this difference-based idea by asking him, "How would you characterize the precision of the *group as a whole*?" After thinking about this for some time, Robert suggested that he would average the differences between the mean and each measurement. Distances corresponding to over-estimates were positive and those corresponding to under-estimates were negative. Robert proposed to find their sum and then to divide by the number in the sample. Robert thought that this method would be "like the average," except that it would indicate how close the measures were, "on average." When he attempted to find the mean of the differences, he was surprised that the sum was zero. (This is a property of the sum of differences between each observation and the mean. It is a consequence of the definition of the mean.) 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 (See Figure 10)

Figure 10

In light of class discussions about some estimates being over and some under the real height of the flagpole, the teacher 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. The teacher built on this student insight to introduce the absolute value function. Robert used the absolute value function in the Tinkerplots™ formula menu to generate the average deviation. He then plotted the absolute values of the differences, and located their average value—the average deviation (see Figure 11).

Figure 11

A variant on this method was appropriated by a pair of fifth-grade students in the current design study who found the differences between each observed value and the median. This method was prompted by their consideration of potentially perfect agreement among the measures. In this ideal case, they suggested that the spread would be indicated by zero. The instructor asked how they might define their measure so that zero would result. The instructor's question bootstrapped students' consideration of difference. Working from this basis, students first obtained sums of the absolute values of the differences. 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. meter stick for arm-span). The instructor 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 (see Figure 12). The instructor’s suggestion that the median difference represented “typical closeness” appeared to stabilize this distinction (meaning that when presenting to classmates, they were able to clearly articulate the distinctions).

Figure 12

Center-clump solutions. A group of students claimed, “where the precision was where most people had their numbers.” Then, they found that 50% of all measurements were in the 40s, and 28% of all measurements were in the 50’s. So, they decided to use the percentage of measures in the decade-interval containing the mean as their measure of precision. The teacher worked further with one of these students to capitalize on the hat plot function of Tinkerplots™ to characterize this notion of center-percentage. First, the student displayed a 25-75 percentile hat plot, like that displayed in Figure 13. Then, the student used the reference line function of TinkerPlots™ to find the values bounding this interval (44.6, 53.6) The student subtracted 44.6 from 53.5 and said that 8.9 was a measure of precision because it captured closeness. When more precise tools were employed, the measure became smaller, in alignment with a “tighter” center clump.

Figure 13

_____

Modeling Error

To date, we have engaged students in modeling error with a random device, a spinner, in only the first design study conducted with a mixed class of fifth- and sixth-grade students. In this study, students measured the height of the school's flagpole.

Sources. Students in this class identified several potential sources of error. One was "wobble," the error introduced by slight variations as one tried to use a tool to line up or sight on the top of the flagpole. Wobble results in prospective errors of angle of measure. For example, at a distance of 50 feet away from the flagpole, a 1 degree error in the angle of the line of sight will produce an error in the height estimate of the flagpole of about 2 feet, resulting in either overestimates (+2 feet) or underestimates (-2 feet). Students did not know trigonometry, so the instructor suggested this magnitude as a way of characterizing "a little wobble." Students mentioned that "medium" and "a lot" of wobble were much less likely. A second source of error suggested by students was due to the contour of the ground. The tape measure employed to find the distance between the base of the flagpole followed the contour of the ground, and this introduced error as students attempted to stretch the tape and otherwise compensate for the effects of contour. Students judged this error as less costly than wobble, and thus suggested magnitudes within a foot as typical errors, with larger errors of 2 or more feet. The third source of error suggested by students was the measure of their height, and it was generally agreed that this error was the least consequential (although some students misunderstood the right triangle

model and thought that shorter people were at a permanent disadvantage!). These were the primary sources suggested by students.

Area models of chance. We introduced students to a ten-region spinner and asked them how we might use the spinner to think about how likely each kind of error would be for a single source. Most students seemed to have little idea about how one might make this mapping, and one denied that it could be considered because the spinner and the measurements were not alike. For this student, we created alternative investigations of the properties of spinners.

One student suggested that there would be a relation between the amount of space on the spinner and the chances of making that kind of error. We followed up on this by asking students to think of the relative proportion of a little wobble, a medium amount of wobble, and a lot of wobble for the ten regions of the spinner. Students felt that most of the time there would be a little wobble, and it would be hard to determine its direction. The instructor suggested that perhaps 60% of the measurements would have this kind of error. Students then worked in small groups or individually to assign regions to spinners. Some students partitioned contiguous regions, others worked under the apparent assumption that partitioning non-contiguous regions would be fairer. A typical solution was to assign 3 regions to a +2, 3 regions to a -2, a one region each to values of -4, +4, -6, and +6.

Combining spinner results. After assigning regions to each magnitude of error to model the chance of making that type of error (one spinner for each source), students worked in teams to collect the results of each spinner for 30 trials. They summed the errors for each trial and added these to the imagined "real" length of the flagpole. We noticed that students were often surprised to find that occasionally the net error was zero, even when none of the simulated errors was zero. They also were surprised to find that unlikely did not mean impossible. A few of the

trials resulted in very large error sums. Even more surprising was that the resulting shape of the distribution was like that obtained when they had actually conducted the measurements. As one sixth-grader put it when describing the results of his simulation (of measures of the apogee of a rocket launched in the last phase of the design study): “So that basically tells that all five of these different kinds of error pretty much don’t do a whole lot to the damage of the original height.” His comment was addressing the central tendency he observed in his simulation, which was like that of the values observed in the collective measurements.

When we asked students to model a new situation, the distribution of measurements of a rocket’s apogee, only two appeared to make substantial progress in constructing spinner models, although all could readily acknowledge different prospective sources of error. Figure 14 displays the results of some of the trials conducted by one of these students, who modeled five sources of error, including new sources of “person” (sometimes you’re more careless) and “wind.” As we mentioned previously, this student was impressed by the simulation’s recovery of the central tendency despite even more sources of error (compared to his first simulation of the flagpole scenario).

Figure 14

Discussion

The conduct of two iterations of this design study, with a third in progress, enables contrast with a previous iteration of the design conducted by Petrosino, Lehrer, & Schauble

(2003). In the previous design study, we worked with students whose teachers were engaged for prolonged periods of time in thoughtful reform of their mathematics and science practice. This institutional capacity resulted in students oriented toward understanding mathematics as a sensible system for thought. Consequentially, this meant that practices that are cornerstones to this learning progression for statistical reasoning, such as inventing measures and representations, and engaging in conversations about their merits, were not entirely foreign to them. In contrast, the three iterations of the design reported today are embedded in an urban institutional structure with less capacity. Teachers are embarking on consideration of their practices but have not yet consolidated their grip on new practices (Knapp, this session). Students tend to be oriented toward mathematics as a calculation (Thompson), and practices such as inventing representations or measures have no place in the epistemology of doing things to numbers. Hence, the contrast between contexts (context here is meant to include histories of mathematical learning) serves as a test-bed for the adaptive character of the design. How robust is the design? What seems to make a consequential difference for learning?

We begin with points of similarity among the first and subsequent iterations of the learning progression, because for us, these constitute conceptual replications. First, the context of repeated measure grounds reasoning about distribution, because students can take the perspective of an agent-measurer. From an agentive perspective, distribution emerges from the parallel activity of many agents measuring the same attribute. Everyone does not get the same measurement, because some agents are more mistake-prone than others (in these accounts, the attribution of error is usually reserved for the other person). However, most agents are pretty "careful," so there are more measurements in the center than elsewhere. Thus, central tendency is a consequence of the measurement process, and the focus of the tendency, the center, emanates

from the invariance of the attribute. As one fifth-grader recently explained, the circumference of a person's head changed as they grow (she had in mind RL's), but not in one day. So she felt safe saying that the circumference of the head measured that day remained unchanged during the course of the measurements. Changes in means of measure, such as changes in methods or tools, have real consequences for agents engaged in measurement. Hence, students anticipate that use of better tools or methods of measure changes the spread of the distribution. In sum, many of the features of the distribution that are readily apparent (after due consideration of representation) have counterparts in process. Distribution results from the outcome of a process and is not just "out there" as a quality of a collection.

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 teachers 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 center and spread support consideration of just what one might mean by each. Thus, there is an intimate relation between conceiving of the "centeredness" 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, Schauble, Carpenter & Penner, 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 meter 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)?

We turn now to bumps in the road, some of which we believe can be attributed to the different histories of mathematical learning between the suburban and urban settings. First, the nature of the measurement affects the transparency of the measurement process for students. When students in the Petrosino et al. (2003) measured the heights of flagpoles, they were familiar with properties of triangles. In contrast, although they had received instruction about triangles, students in our first two design studies were not. Hence, the measurement model was less transparent, and some students in the first design study in the urban school thought that shorter people were more likely to make errors. The source of this misunderstanding is not clear to us, but more than one student voiced it. Other sources of error were transparent to the students (e.g., wobble), so this misconception was more unsettling than fatal. The remedy adopted in the second iteration of the design study was to include instruction about triangles, so that the measurement model would be more sensible. But this instuction took much instructional time,

and although we valued what students learned, others might not be so enthusiastic. The remedy in the third design study was to switch to direct measurement. Students measured the length of their teacher's arm-span, the circumference of a researcher's head, and the area of another researcher's handprint. Because direct measure is also not transparent, these activities rested on a base of prior lessons in measurement (see Lehrer, Jaslow, & Curtis, 2003) designed to support the growth and development of a children's theory of measurement.

Second, the pace of instruction has been much slower in the urban setting, because we could not rely on classroom norms of mathematical explanation. Students were not used to taking other students' work into account, nor, as we suggested previously, were they comfortable with the very idea that they could invent--anything. Statistics were experienced as received, not as solutions to problems of measure, and representations were conventions found in books. We know of no remedy for this, and in fact, would deplore such a remedy. But, because education is a normative enterprise, we can foresee many classrooms in which the practices we advocate would either not be welcome or would be appropriated in a manner that would result in lethal mutation.

We conclude with a brief exploration of issues that we have encountered that multiple instantiations have helped make more visible to us. First, we are interested in understanding better how students might come to model chance errors of measurement. There are potentially many obstacles, which include establishing the relation between area and chance for a spinner, considering different sources of error, and conceptualizing long-term, repeated random processes. However, our sense is that all of these challenges are less than that of establishing the mapping between the phenomenon and the model. In related work, we have found it advantageous to ground model-world mappings in resemblance, with gradual lifting from literal

resemblance to representational systems (Lehrer & Schauble, 2005; in press). Modeling chance in the context of measurement error imposes an additional burden: The relation between the representing and represented worlds is syntactical, only. As one of our students stated plainly, there is no resemblance—at all! Perhaps syntactic models will prove beyond the ready grasp of most students of this age.

Second, although we have not addressed this issue here, we are exploring students' conceptions of sample-to-sample variation under the guise of predicting what might happen if we measured again. So far, we have spent more time with students' conceptions of directed variation (e.g., changes in methods, tools) than with undirected, random variation. We are uncertain about how to build on students' intuitions in mathematically productive ways in light of their difficulties modeling error.

Third, as in any design study, there are a series of local contingencies that point toward the need to examine conceptual development in greater detail, to conduct conceptual analyses that were initially neglected. For example, as we mentioned earlier, children's conceptions of splitting the data in half instigated a contrast between distance- and count-based senses. The former corresponds to the midpoint of the range and the latter to the median. We found the contrast for children less transparent than we expected and so modified the design on the spot.

Fourth, the current work with the fifth-graders suggests a pathway for chance that paradoxically begins with determinate conceptions. Many students initially attribute distribution to mistakes made by measurers. Mistakes are considered volitional. If only the measurer had been more careful. For example, one of the fifth grade participants in the current study predicted that other measurers outside of the classroom would produce a much more scattered set of measurements, because they had not had the advantages of participating in the class. "They" did

not know that you had to consider the nature of units of measure, nor were they aware of the virtues of uniform methods for measuring, etc. Volition was challenged by experiences of being very careful yet still not completing agreeing about the measurements. No matter how hard one tried, mistakes still occurred. We have come to think of this as a pathway to chance, as children re-consider what at first appears completely determinate.

References

Chi, M. T. H. (2005). Commonsense conceptions of emergent processes: Why some misconceptions are robust. *The Journal of the Learning Sciences*, *14*, 161-199.

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

Horvath, J. K., & Lehrer, R. (1998). A model-based perspective on the development of children's understanding of chance and uncertainty. In S. P. LaJoie (Ed.), *Reflections on statistics: agendas for learning, teaching and assessment in K-12* (pp. 121-148). Mahwah, NJ: Lawrence Elrbaum Associates.

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.

Liu, Y., & Thompson, P. W. (2002). Randomness: Rethinking the foundations of probability. *Proceedings of the Twenty-fourth Annual Meeting of the International Group for the Psychology of Mathematics Education*.

Petrosino, A., Lehrer, R., & Schauble, L. (2003). Structuring error and experimental variation as distribution in the fourth grade. *Mathematical Thinking and Learning*, *5* (2&3), 131-156.

Saldanha, L. A., & Thompson, P. W. (in press). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics*.

Thompson, P. W. (2000). Radical constructivism: Reflections and directions. In L. P. Steffe & P. W. Thompson (Eds.), *Radical constructivism in action: Building on the pioneering work of Ernst von Glasersfeld* (pp. 412-448). London: Falmer Press.

Figure1. Construct: Data display

Construct: Data Displays

Level	Performances	Example
2.7 Integrated case-density perspective.	2.7.3. Extend interpretations beyond the research question	"The distribution of the data is wider for rounded-nosecone rockets than for pointed-nosecone rockets. Maybe that's because pointed rockets flights are more consistent.
	2.7.2. Relate density-view to qualities of case values, and vice versa	Relate qualities of a case as an example of general qualities of region of data (case as typical of data region)
	2.7.1. Make generalizations about display features	"If the interval on the display is the same as the range of the data, then all the values will be stacked in one bin."
2.6 Density-based perspective	2.6.3. Create different visualizations of density-based displays based on their utilities	"In this graph the bins are small, so there are many holes in the graph and you can't see the clumps very well. In the other graph, the bins are bigger, so the clumps are more clear to see."
	2.6.2. Explain how qualities of the density-based display are consequential for answering the research question	"I found out that measurements between 45 and 55 were 70% of our measurements. So, I guess the true height is somewhere between 45 and 55."
	2.6.1 Take the resulting distribution of data display into consideration when deciding on bin size	Talk about relationship between shape of data and choice of interval
2.5 Emerging density-based view to the data	2.5.1 Manipulate bin size in display, but do not coordinate bin manipulations with questions.	Says that a larger bin will result in fewer number of bins, and very likely more cases in each bin, but has no way of deciding whether some bin size is better than another
2.4 Aggregate-based perspective	2.4.2 Create bin displays with appropriate axes	• Use continuous scale for continuous data • Create display with appropriate order, for example, small, middle, large, or large, middle, small • Use equal intervals (so holes, if any, are visible)
	2.4.1 Create frequency displays highlighting repeated values with appropriate scale	"Number line" display
2.3 Emerging aggregate view of data	2.3.2 Create bin display to show groups of similar values (but bins may not be of the same size or bins may be juxtaposed without regard to interval)	Create unordered bins, and commenting on, for example, the number of 40s vs. the number of 50s. Put discontinuous (and unequal interval) bin names when asked to fill in bin name with a provided display and corresponding data set (e.g., a display with an interval of 20): e.g., 2-15, 25-36
	2.3.1 Note similar values or "clumps" in the data set.	Notice "plateaus" in the case display or a group of similar values.
2.2 Case-based perspective	2.2.2. Identify significant individual data points, not necessarily considering their role in the whole data set.	Identify outliers, maximum, and minimum.
	2.2.1. Order data.	Order the set of values: (45, 46, 51, 52), but do not attend to interval properties (if any). For example, fail to notice that the gap between 46 - 51 is bigger than the gap between 45 – 46 or 51 – 52 Order all data on the display and stack repeated values, without showing holes.
2.1 See data as a collection of numbers	2.1.2 Manipulate data in a way that suggests losing sight of the goals of the inquiry. Activity appears guided by "doing things with numbers" than by answering a research question.	Group numbers based on whether they are odd or even. "We grouped every 5 values. So 31, 33, 35, 42, 43 are in one group and 44, 44, 45, 46, 47 are in one group." When asked why, students smile and shrug.
	2.1.1. Read values but do not attempt to group, order, or do any other manipulations on them. Don't see data as alterable or as subject to manipulation.	"Everyone got a different value!"

Figure 2. Simply ordering data

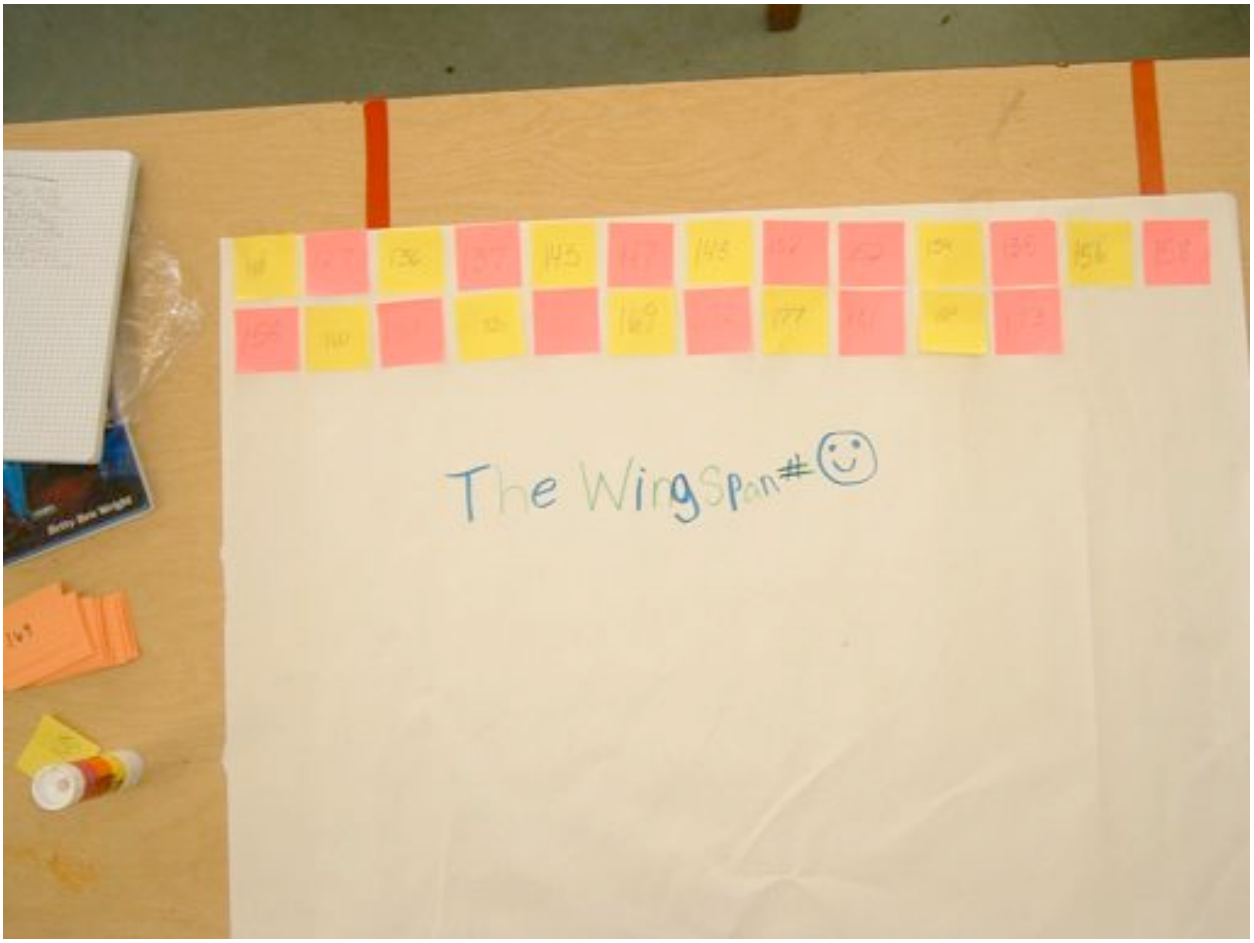

Figure 3. Ordered value display

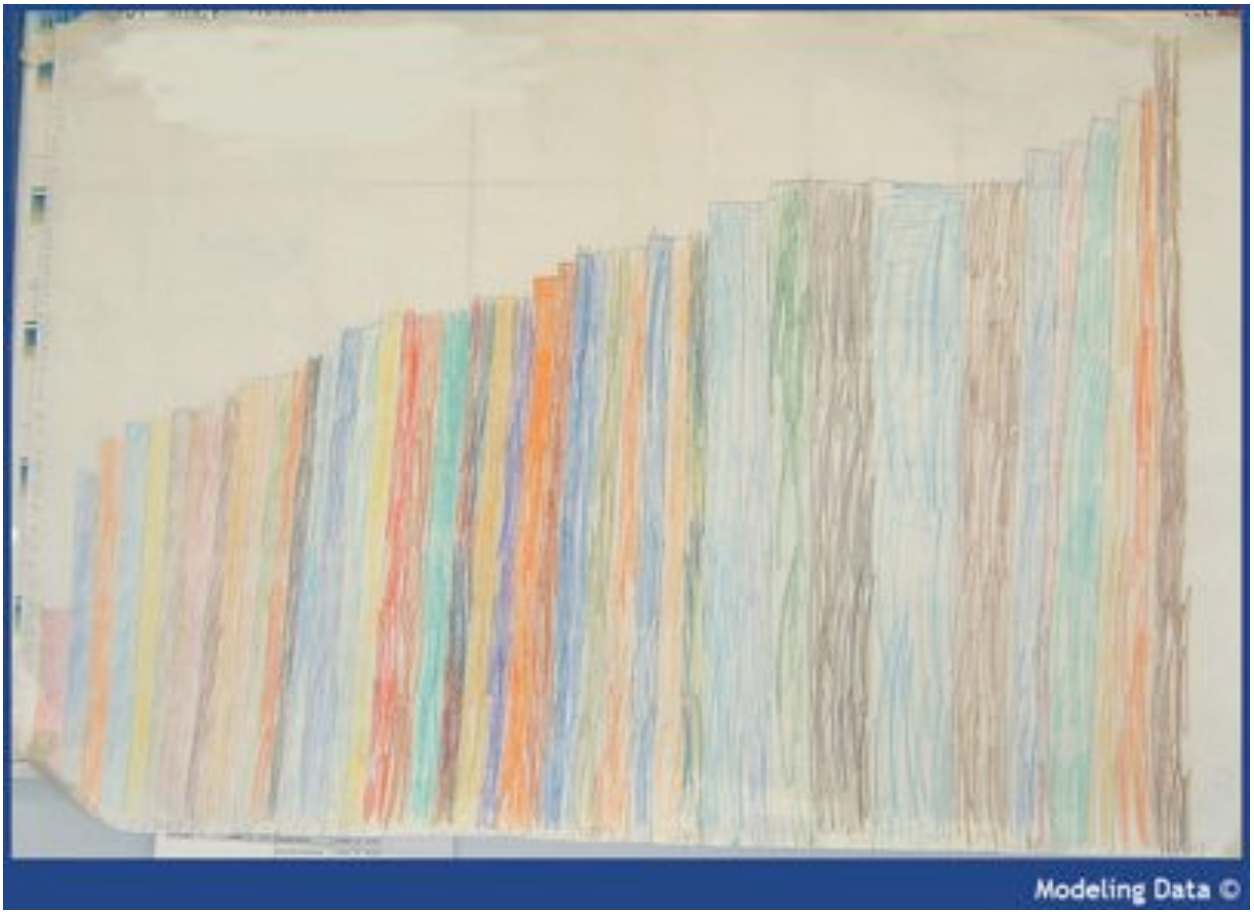

Figure 4. Ascending values

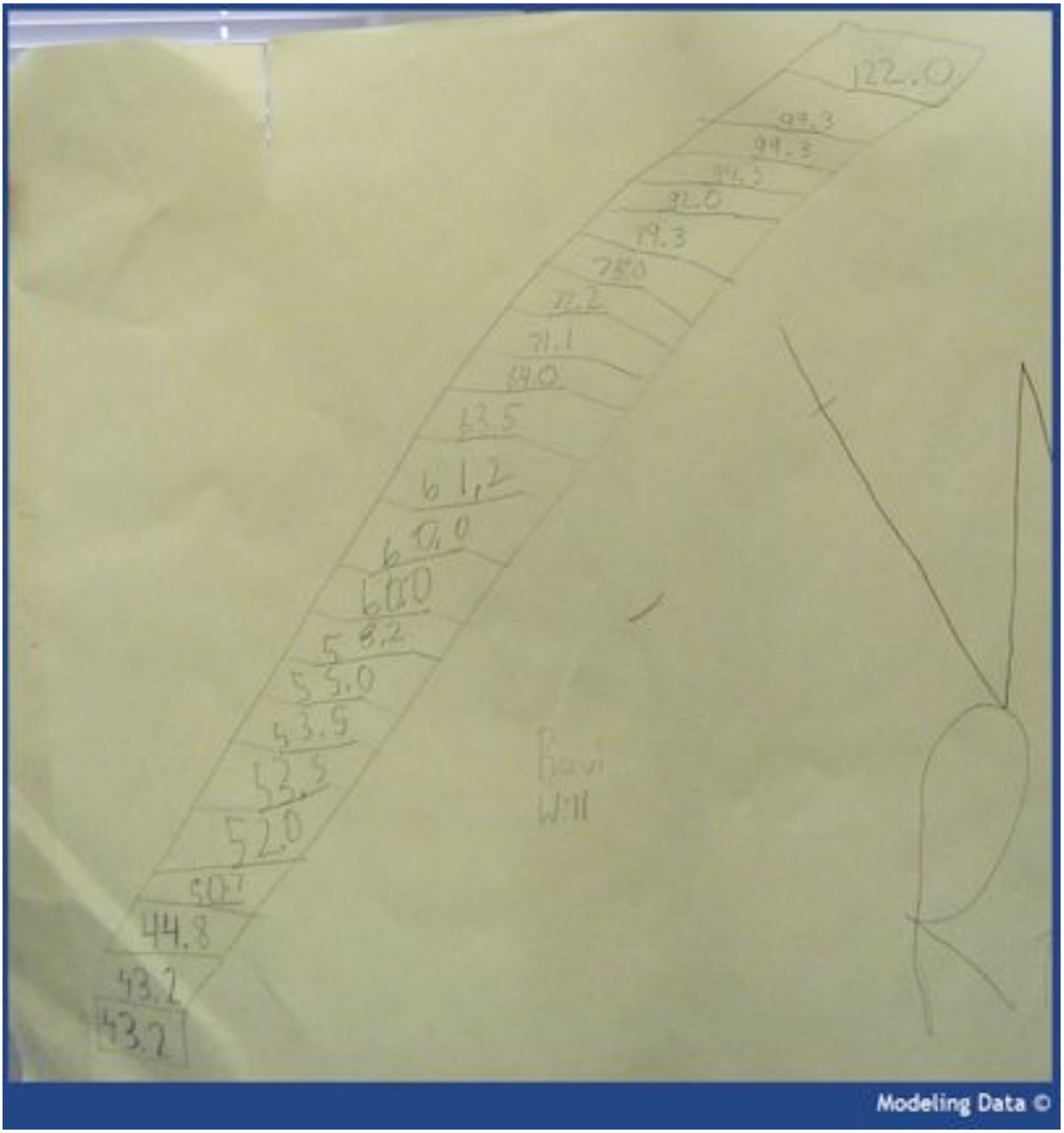

Figure 5. Ordered case frequency display

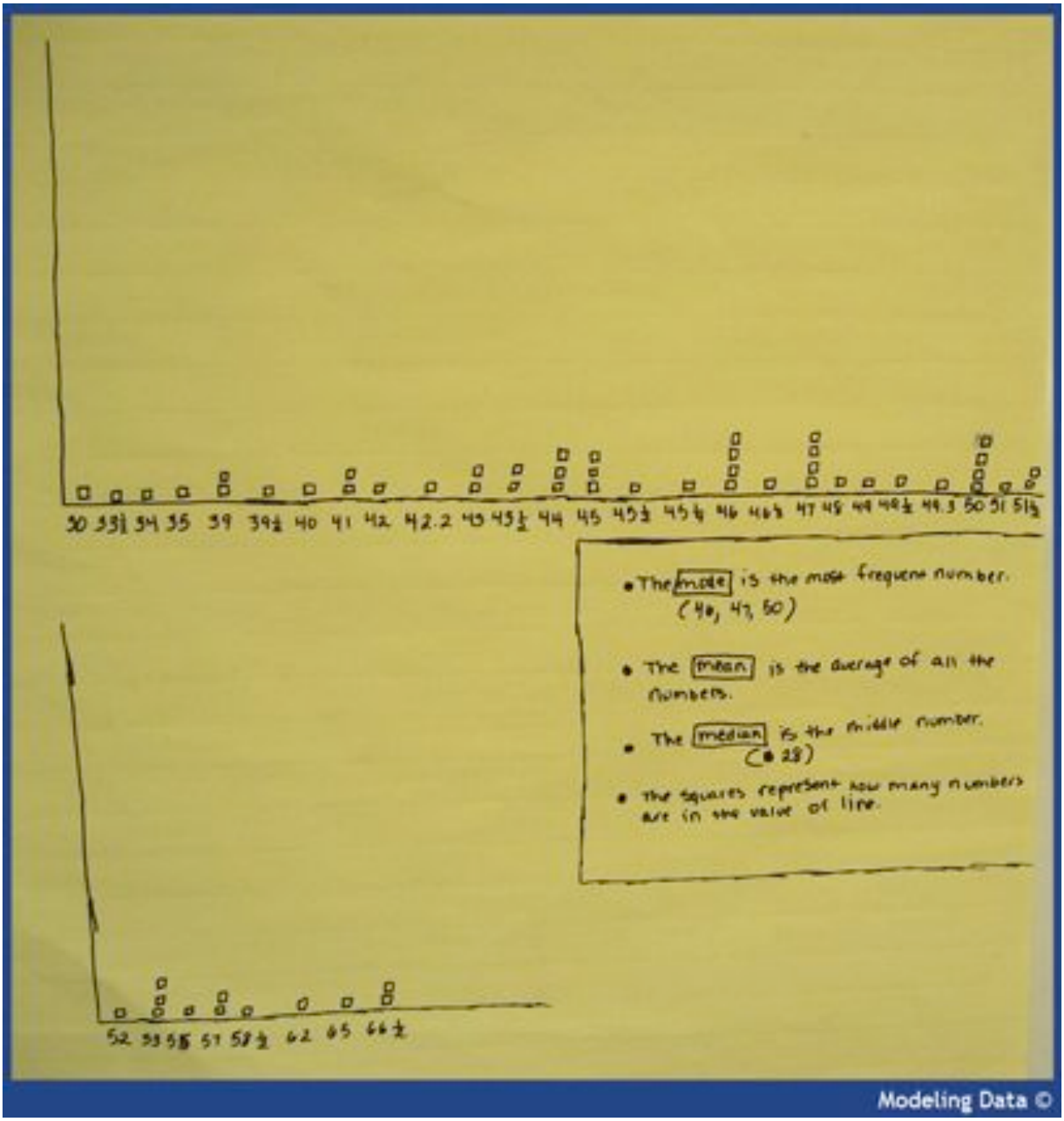

Figure 6. What's missing?

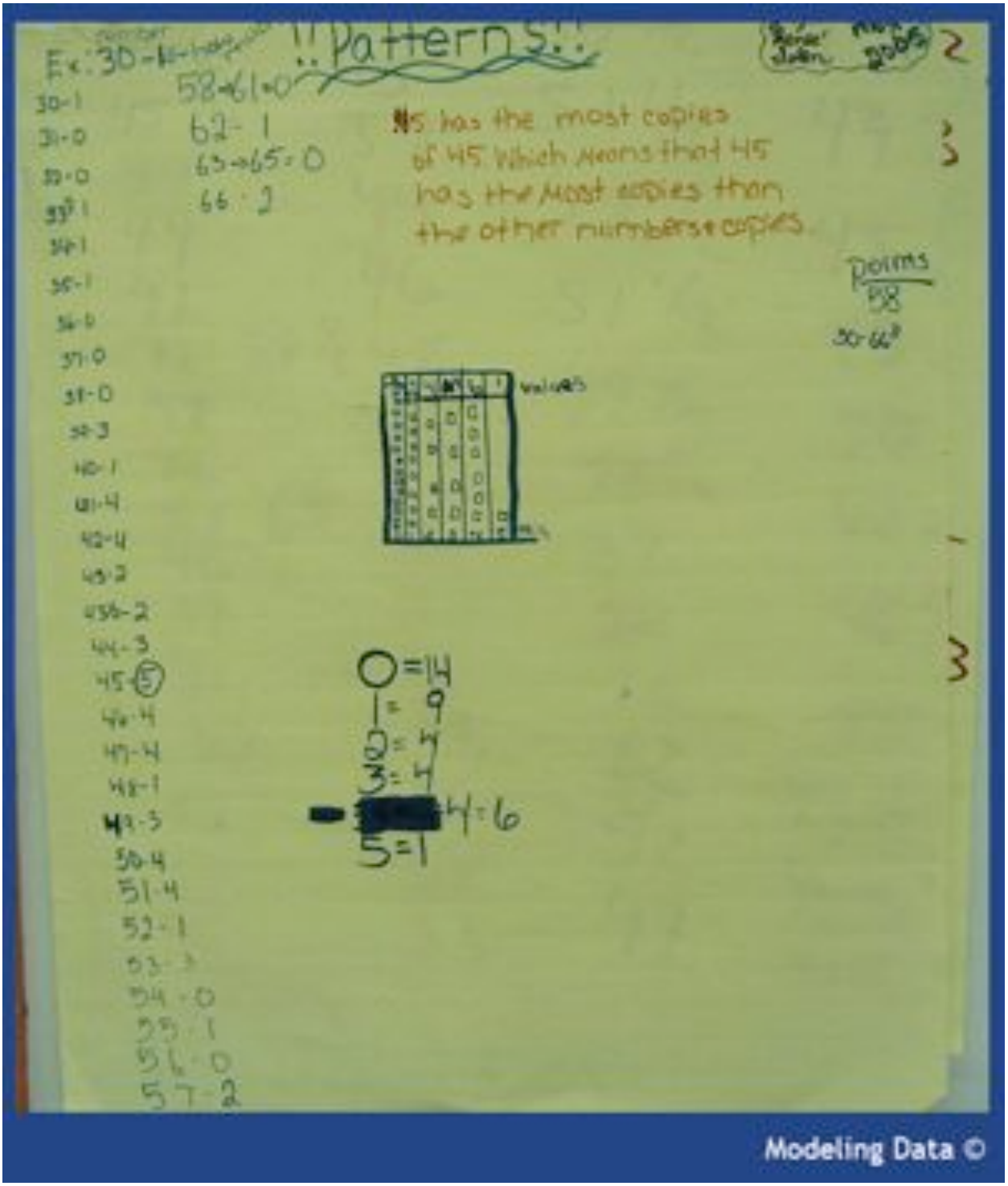

Figure 7. Case frequency display with missing values

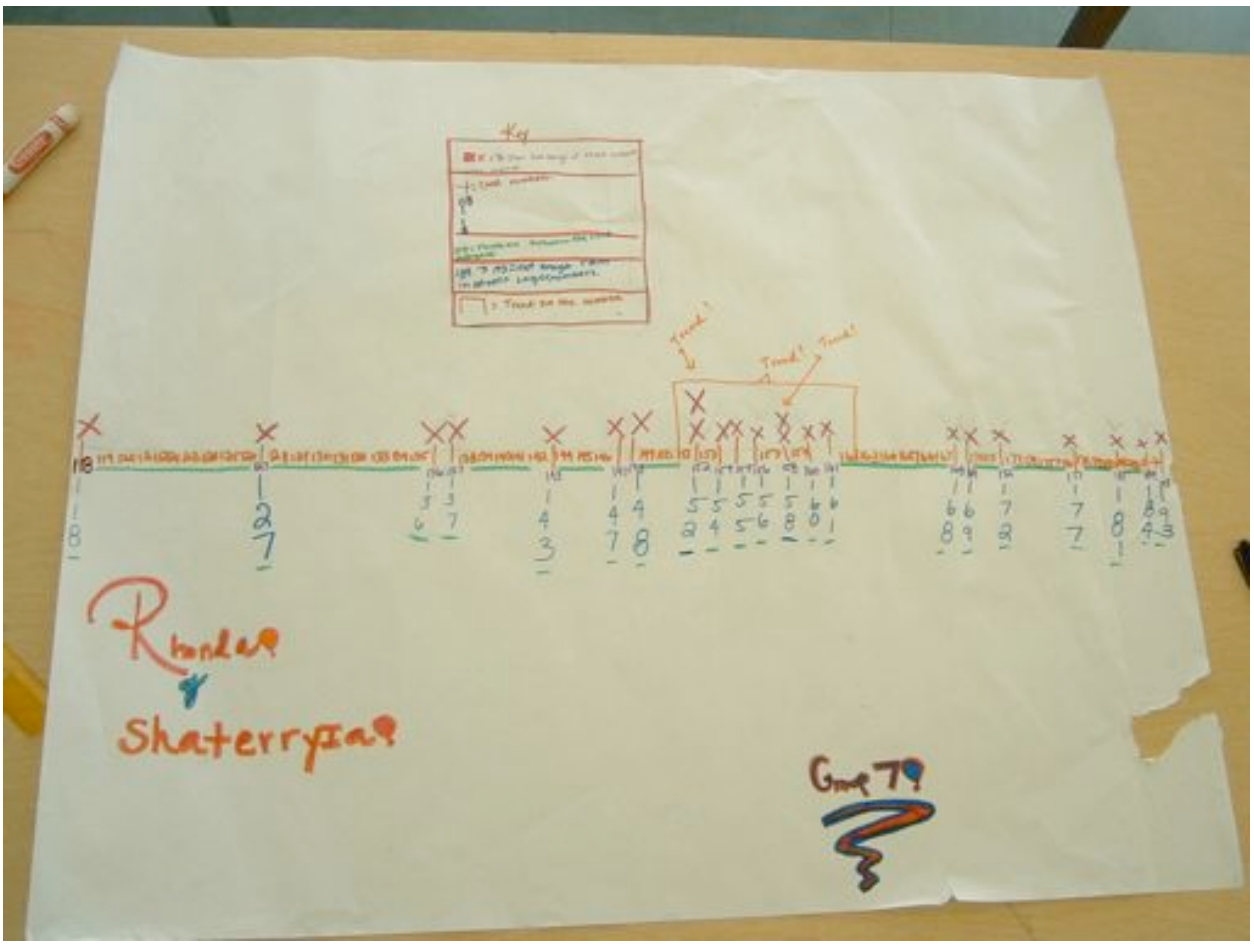

Figure 8. Ordered bin display

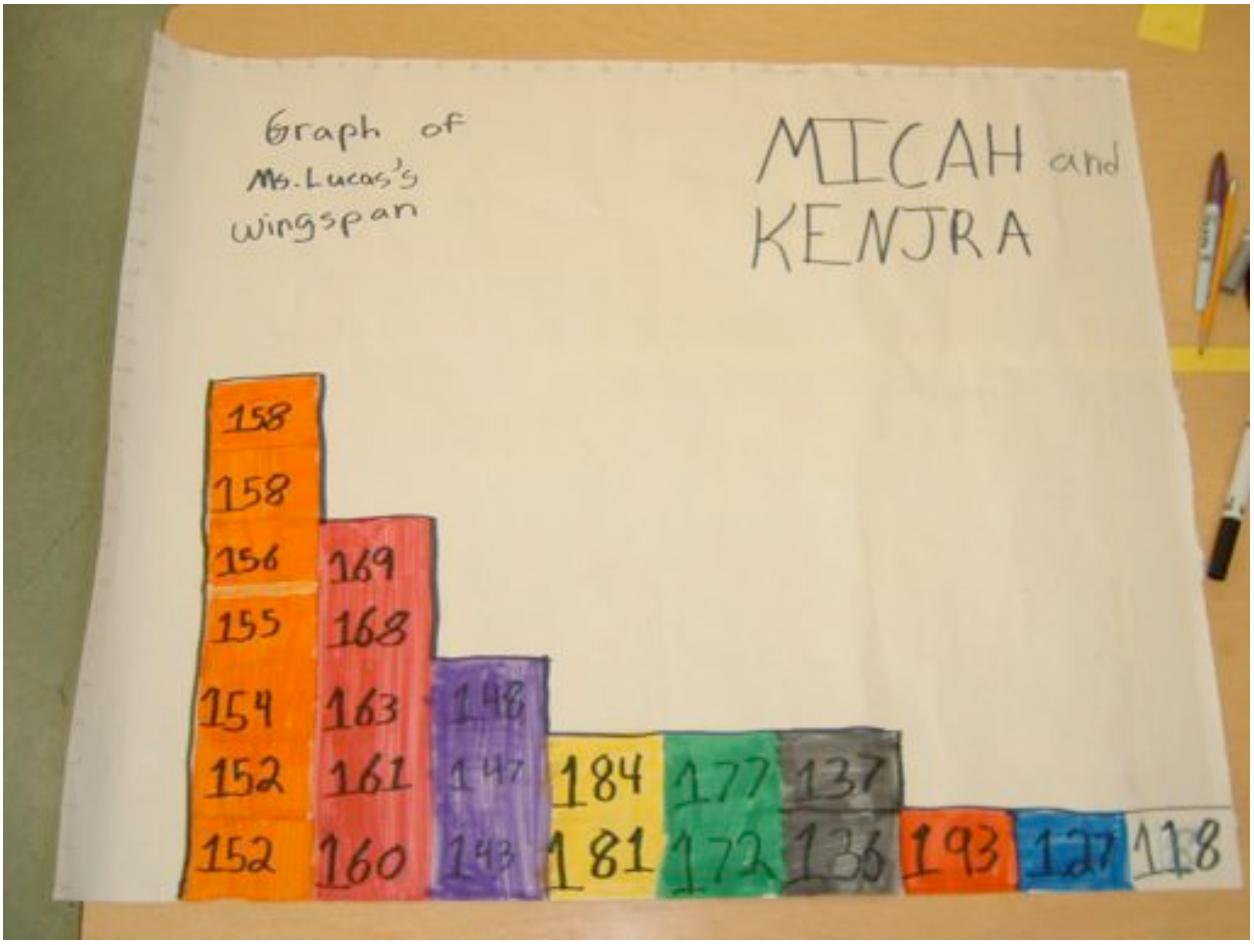
Graph of
Ms. Lucas's
wingspan
MICAH and
KENJRA
158
158
156
155
154
152
152
169
168
163
161
160
148
147
143
184
181
177
172
137
136
193
127
118

Figure 9. No pattern!

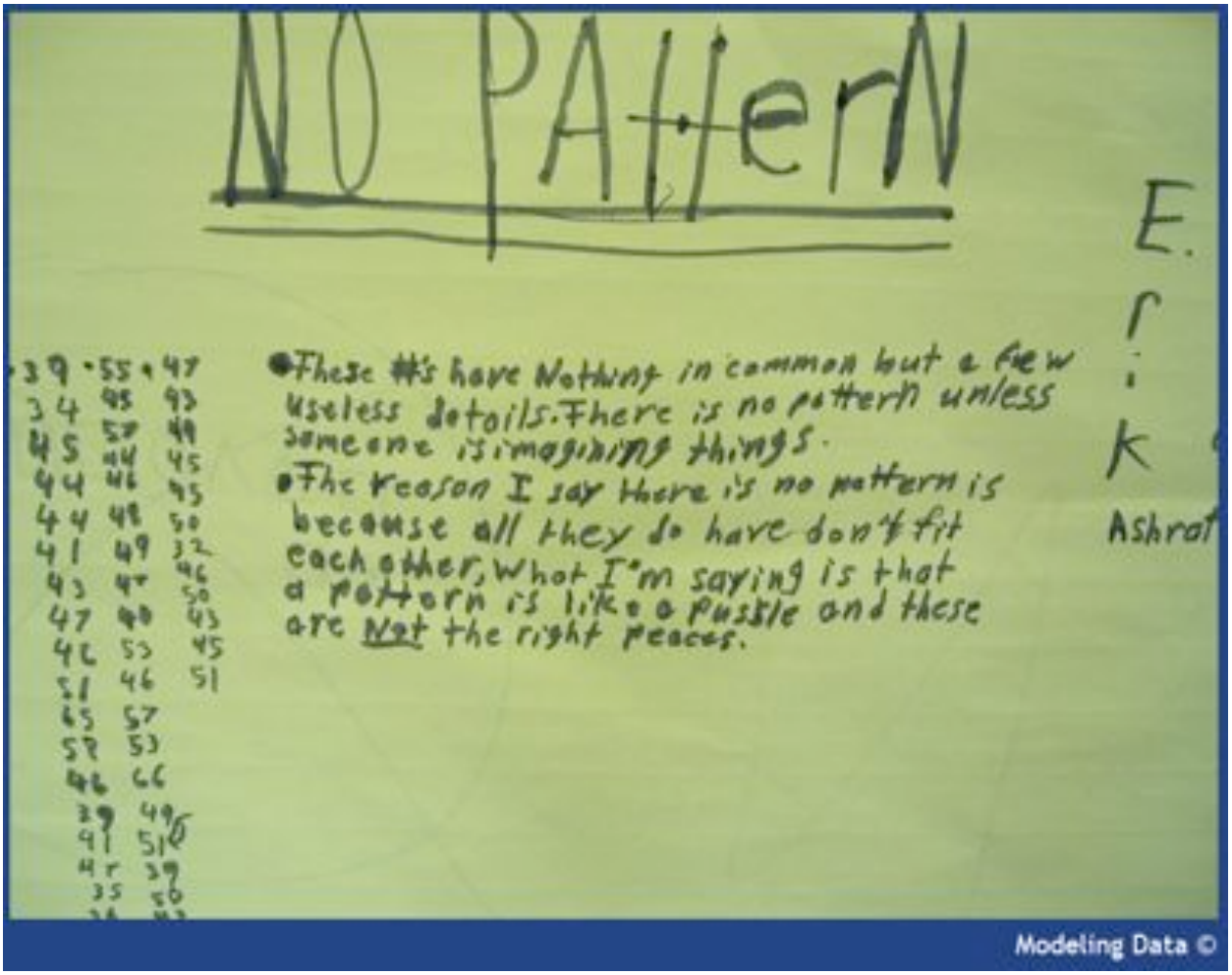

Figure 10. Tinkerplots™ *plot of signed differences*

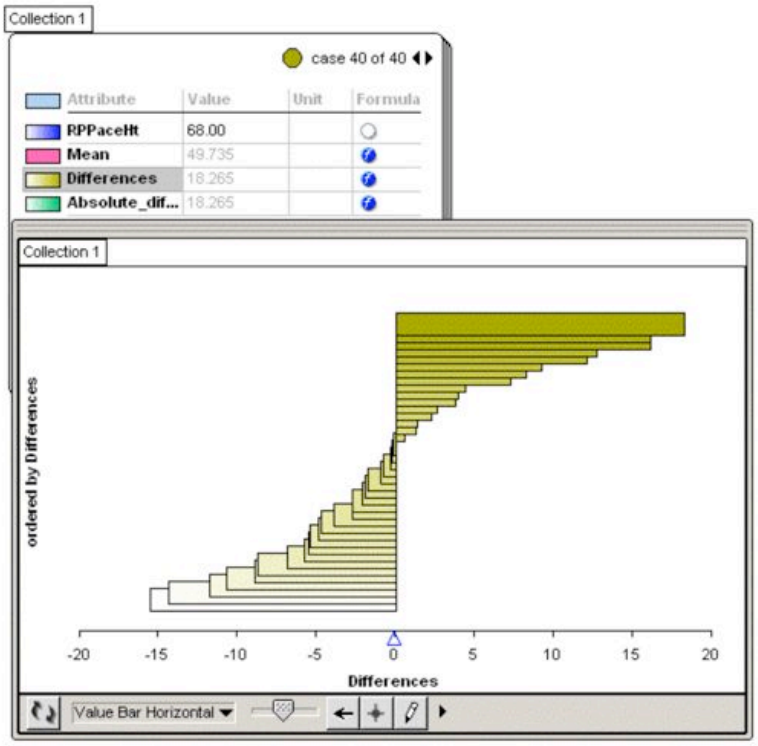

Figure 11. Tinkerplots™ *plot of absolute values of differences and average deviation*

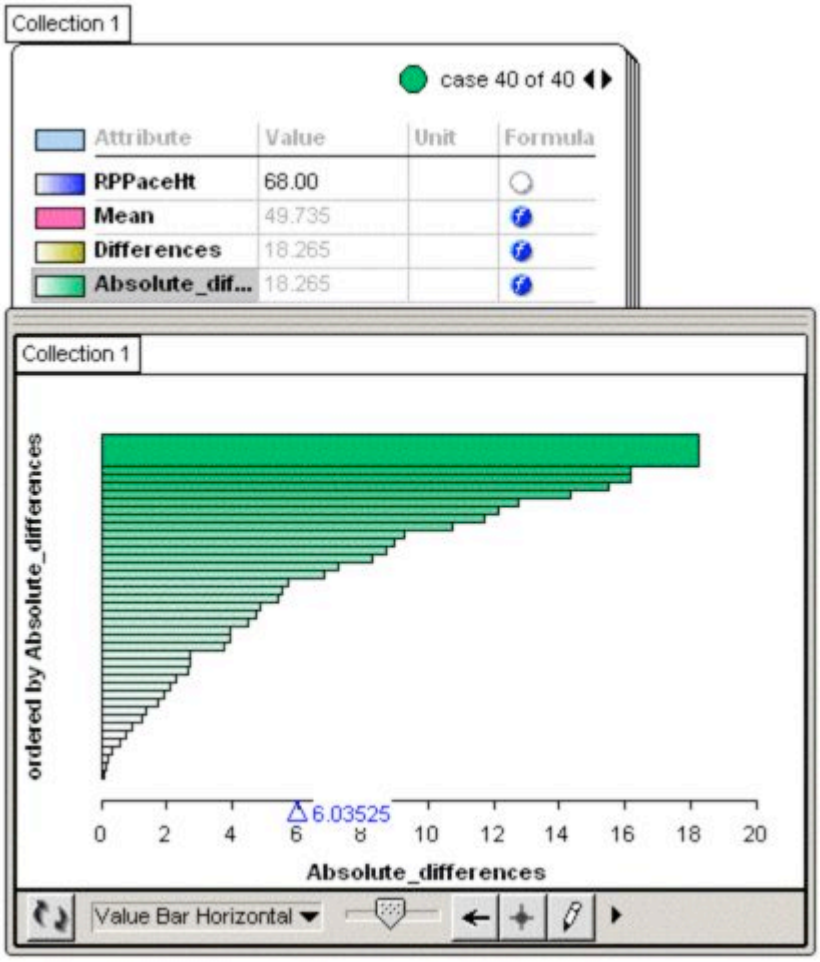

Figure 12. Tinkerplots™ *plot of absolute values of differences and median of the absolute differences*

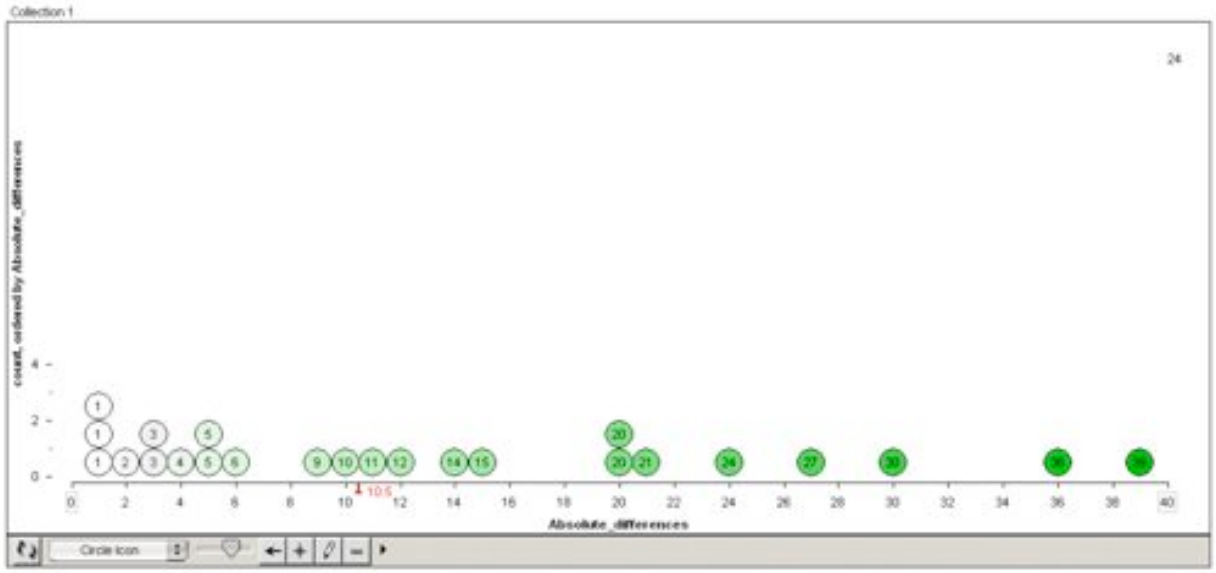

Figure 13. *A 25-75 percentile hat plot*

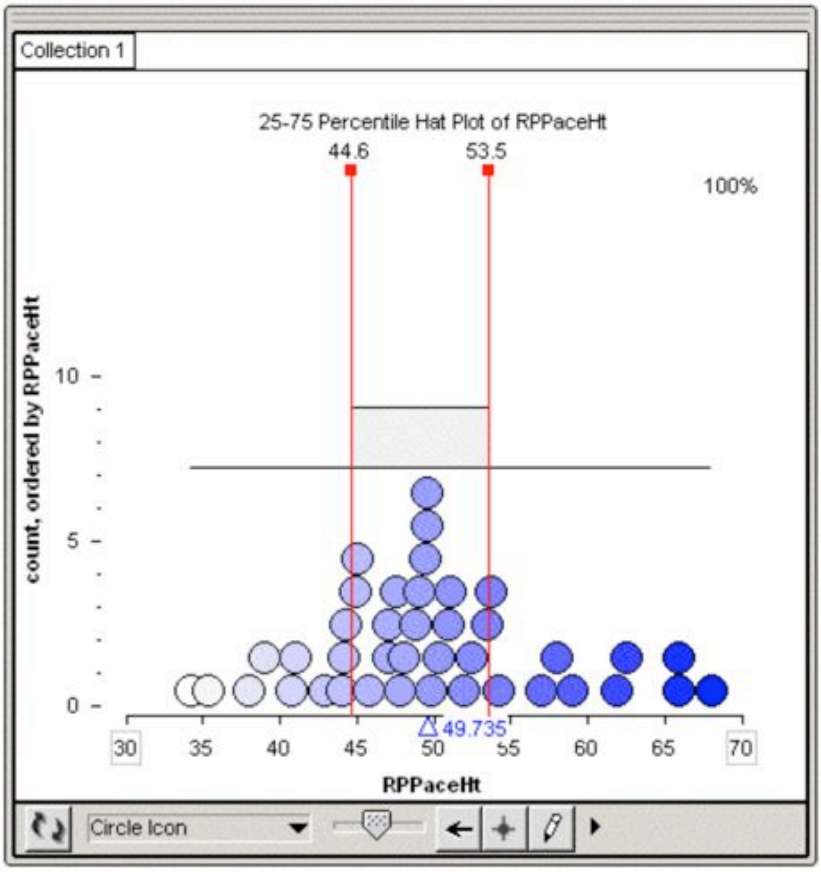

Figure 14. Modeling sources of errors with spinners