



## Getting a Grip on Variability

Richard Lehrer<sup>1</sup>  · Leona Schauble<sup>1</sup> · Panchompoo Wisittanawat<sup>1</sup>

Received: 27 March 2020 / Accepted: 17 July 2020  
© Society for Mathematical Biology 2020

### Abstract

Because science is a modeling enterprise, a key question for educators is: What kind of repertoire can initiate students into the practice of generating, revising, and critiquing models of the natural world? Based on our 20 years of work with teachers and students, we nominate variability as a set of connected key ideas that bridge mathematics and science and are fundamental for equipping youngsters for the posing and pursuit of questions about science. Accordingly, we describe a sequence for helping young students begin to reason productively about variability. Students first participate in random processes, such as repeated measure of a person's outstretched arms, that generate variable outcomes. Importantly, these processes have readily discernable sources of variability, so that relations between alterations in processes and changes in the collection of outcomes can be easily established and interpreted by young students. Following these initial steps, students invent and critique ways of visualizing and measuring distributions of the outcomes of these processes. Visualization and measure of variability are then employed as conceptual supports for modeling chance variation in components of the processes. Ultimately, students reimagine samples and inference in ways that support reasoning about variability in natural systems.

**Keywords** Education · Modeling variability · Random processes · Sampling variability · Informal inference

Given the diversity of the sciences and the many goals and forms of scientific practice, science education needs to be driven by a coherent vision of what is most worthwhile for students to learn. Moreover, that vision not only should encompass ultimate goals, but must also include an account of how instruction from the earliest grades can provide a strong foundation for ideas that may not come to fruition until years later. We start from the premise that science is inherently a modeling enterprise (Nersessian 2008; National Research Council 2012; Windschitl et al. 2008), a premise that leads to the question: What kind of conceptual repertoire best initiates students into the modeling game (Hestenes 1982)? We have been pursuing this question for the past 20 years in

---

✉ Richard Lehrer  
rich.lehrer@vanderbilt.edu

<sup>1</sup> Vanderbilt University, Nashville, USA

partnership with participating school districts in three different states. Elsewhere we describe how we introduce children to inventing and revising models as conceptual means for understanding the functioning of natural systems (Lehrer and Schauble 2019). Here, we focus attention on a set of conceptual core ideas that have proven to be especially powerful in children's developing modeling repertoire. In particular, we seek to describe and illustrate how variability and the associated idea of uncertainty can equip even young students to more deeply understand the world.

Characterizing and accounting for the variability of a system are often consequential for theory development and model building, especially in the biological sciences. For instance, evolutionary accounts of diversity require a firm grasp of distinctions between directed and random variation in populations of organisms. Understanding many ecosystem processes is accomplished by sampling in space and time, which, in addition to practical matters of instrumentation, measures, and methods, is also distinguished by variability, both within and between samples (Lehrer and Schauble 2017). Consequently, interpretations based on samples should account for sampling variability. Understanding random sampling variability hinges on a hierarchical image of sample in which any sample is simultaneously composed of outcomes that vary and is also a member of an imagined collection that varies from sample to sample (Thompson et al. 2007). Of course, these ideas are prominent in other scientific domains as well and are key in both observational and experimental methods. The ubiquity and prominence of variability in scientific inquiry suggest that students should get some mental grip on the challenges of variability, such as a hierarchical image of sample, during the course of schooling. Yet, developing a conceptual and practical grasp of variability is often challenging, even for students at university level.

How, then, is it possible to provide comparatively young students with an entrée to some of the conceptual foundations needed to reason productively about variability? We illustrate supports for learning that emphasize student participation in visualizing, measuring, and modeling variability. These practices are accessible to students throughout schooling, and they can provide a good foundation for developing conceptions of variability and uncertainty. To illustrate what this kind of instruction looks like and to explicate the ideas about variability that students can grasp, we describe a sequence of instruction with grade 5/6 (ages 10–11) students and report the typical forms of student reasoning that we observed. First, we outline theoretical orientations that guided the design of instruction. Then, we explain how we staged student participation in practices of visualizing and measuring variability and describe what students tended to learn about variability by participating in these practices. After doing so, we zoom in on the main empirical focus of this report, demonstrating how students' invention, revision, and contest of models of random variability helped them develop more disciplined views of sample and inference. As we will describe, our inquiry centers on how students tended to reason about inference in light of uncertainty, including their image of sample and their aesthetic preferences related to forms of modeling.

## 1 Instructional Design

Our overall goal was to help comparatively young students encounter some of the epistemic issues involved in professional inquiry about and warrant of claims about variability, especially under conditions of uncertainty (Pfannkuch and Wild 2000). For instance, how do particular mathematical choices about ways to construct data visualization result in different shapes of the data? In light of sample-to-sample variation, what does a difference between samples suggest about a claim? The design of instruction was informed by guiding principles (Bakker 2018) that influenced our generation and selection of tasks, tools, and formats of classroom activity. However, guiding principles are not scripts. As in any instructional endeavor, our choices were also influenced by the accessibility of materials and tools in everyday classrooms, teacher preferences, and related contingencies that often could not be predicted in advance, and therefore, were generated during implementation.

### *Learning by Participating in Approximations of Professional Practice*

The most fundamental principle that guides our instruction is that students should receive opportunities to be inducted into forms of activity that STEM fields employ to generate, contest, and revise knowledge (National Research Council 2012). Modeling renders student thinking public, visible, and testable and, hence, subject to critique and revision by other members of the community. Students also experience that even when models are provisionally accepted, they remain open-ended; today's solutions become grist for tomorrow's challenges. When students become engaged in the invention and critique of models, they experience a tension that also confronts scientific professionals—a tension between individual and collective work (Ford 2015; Ford and Forman 2006). That is, individuals innovate in anticipation of critique by the collective. The intelligibility of the work that individuals conduct depends on their ability to align their activity with the forms of activity taken as normative by other members of the community who are trying to accomplish similar goals. Professional critique relies on collective consensus, however imperfectly achieved. And critique itself is not static, either—occasionally, what individuals invent will reset the norms for what the community takes as valued.

This perspective on the development of knowledge demands a classroom ecology that has particular characteristics. We do not expect students to copy what professionals do, because they do not live in professional cultures. Nonetheless, there are key aspects of disciplinary epistemology that we consider critical to carry over into classrooms. First, students should have firsthand experience with the productive tension between innovation and critique. Inventing models helps students directly experience some of the challenges that practitioners confront. Critiquing, on the other hand, confronts students with other perspectives on those challenges. The result might be revisions to one's model, one's reasoning, or even the norms one accepts for appropriate forms of activity. The models that the field considers conventional or canonical can come to be understood by students as perspectives that others in the discipline eventually developed as they confronted challenges like those the students experienced (Jones et al. 2017).

Second, the duration and nature of participation must be sufficient so that students grasp the relationships among the apparently discrete activities they are pursuing. Another way of saying this is that students must come to see how the particulars of their activities make sense cumulatively in light of their overarching goal of generating and critiquing knowledge. Often, science involves the conduct of extended chains of reasoning that are constructed over time as one engages in related activities, such as refining questions, designing measures, collecting data, displaying it, and drawing interpretations. Yet, students do not automatically grasp these relations as activities are encountered. They need sufficient time and support to come to see how otherwise distinct performances are blended to accomplish greater conceptual articulation (Lehrer and English 2018). For example, as we will see shortly, students' initial practices of visualizing and measuring variability can later be employed as conceptual tools for developing and testing models of variability.

These concerns guide the approach that our participating teachers adopt in their instruction, and this approach to instruction will become visible as we begin to describe the conditions that we collectively arranged as students were introduced to ideas about variability. First, however, we briefly describe the sequence of concepts as students encountered them, to provide a sense of how later ideas built on earlier experiences.

#### *Participating in Variability-Generating Processes*

Instruction about variability often leads with preselected samples or involves students with samples of natural variability, such as the heights of students in their classroom. Yet, statistical conceptions characterize samples as representatives of population processes. Treatments of probability also rely on an image of a long-term, random process. For example, a relative frequency of a target event is considered to inform the probability of its occurrence over repeated trials. Unfortunately, students typically understand probabilities in terms of specific, single-occasion outcomes. For instance, an estimate of the probability of rainfall on a particular day is interpreted as meaning that it will or will not rain that day (Konold 1989). Our focus in instruction, therefore, is to help students understand that statistical approaches to sample and chance rely on assumptions about long-term process (Saldanha and Thompson 2002, 2014). This is a challenging concept that typically eludes students, partly because it is based on reading into a collection of cases to conceive of the collection as a sample. For this reason, instructional design needs to make stochastic process visible and tangible to students, so that a sample comes to be regarded as one scoop taken from a river of ongoing outcomes. To make this conception more concrete, students should experience how a sample emerges from a repeated process.

We have learned that an effective way of providing students with entrée to the generation of random outcomes is to begin by exposing them to processes that are characterized by clear signal and readily interpreted sources of noise (error). Contexts in which students can readily propose causes associated with signal and noise help students conceive of an outcome as composed of the contribution of each (Konold and Lehrer 2008; Petrosino et al. 2003). This helps students envision how distribution, a collective property, can arise from the repetition of a process (Konold and Lehrer *op cit.*). As we will describe shortly, we began instruction first by enacting processes in which both signal and noise were readily interpretable. Experiences with these

processes, coupled with investigation of the behavior of random devices, served as a bridge to natural variability, which students typically find more difficult to understand (Lehrer and Schauble 2007). Having outlined this general framework, we next turn to sixth graders and describe how they typically progress as they learn to use statistical practices in modeling contexts.

## 2 Visualizing Variability

Each student is asked to measure the same attribute, such as the arm-span of the teacher, first with a 15-cm ruler and then again, with a meter stick. In this process of repeated measurement, the signal is the actual length of the teacher's arm-span and readily distinguishable sources of error include small slips such as gaps or overlaps that occur as the student iterates the ruler and meter stick. Each student produces a measure with both tools. (To maintain the assumption of independence of trials, we ensure that students do not yet know the value found by other measurers as they conduct their own measures). To the surprise of the students, the collection of measures that emerges from different individual repetitions of the measurement process has a new property—variability. Sources of error, such as those due to iteration, help students account for the variability they observe, and the tendency of the measurements to cluster roughly symmetrically about a center is readily attributed to a fixed portion of the process—the “true” measure of the length. A switch in process—when the same measurers use a meter stick—results in less variability, which can be readily attributed to less error in iteration. A critical feature of this context is that changes in variability are easily interpreted as causal.

To make these ideas visible to students, we encourage them to work together in small groups to invent displays of the class collection of measures of the teacher's arm-span (Petrosino et al. 2003). Directions are deliberately left ambiguous. Students are told that their visualizations should show “some pattern or something that you notice about the measurement,” and the sole constraint mentioned to children is that the display should be constructed to communicate, for example, “...so that someone else just walking into the classroom could look at your display and see what you noticed.”

By this point in their schooling, students typically have already experienced canonical forms of display, such as bar charts and line plots, but we find that these are rarely invoked. Instead, the inventions of small groups reliably reflect a range of orientations to displaying the collection of measurements (Konold et al. 2015). Some students tend to be oriented to cases, as reflected by displays that highlight individual cases, such as those displayed in Fig. 1a, b. In Fig. 1a, students have displayed the collection as an ordered list, and they took advantage of the plane to create a curve. In Fig. 1b, each case value is ordered and represented as a length. Although the measurements are ordered in magnitude, students who focus on case values often neglect to order the measures.

Other students tend to classify data, generating intervals of values, as in Fig. 1c. In Fig. 1c, the intervals are ordered by frequency, although this is tacit in the display.



without values are not visible. Many students tend to leave out these gaps in the data, assuming that lack of data means there is no information to communicate.

Finally, some students tend to reason about the aggregate and employ the measurement scale along with frequency to display the data. These solutions less frequently occur (Lehrer et al. 2014). Other, more whimsical strategies also sometimes appear, usually because students are deliberately trying to create artistically unique constructions. However, we find that the overwhelming majority of students' invented displays reflect the range that we described, from case value to aggregate.

The student invention process is followed by whole-class critique. A student who was not the inventor of one of the displays uses that display to infer what the inventor seemed to notice about the collection of measured values. This is referred to as deciding "what the display shows and hides" about the measurements. Conversations like these provoke students to notice the aspects that are highlighted or backgrounded by the display in question and invite discussion about how mathematical concepts such as count (frequency), order, interval, and scale result in particular visualizations of the data. Teachers seek to help students understand that the shape and other visual qualities of the display are not inevitably inherent in the data, but rather, result from choices that designers make, especially the mathematical concepts they employ to highlight some characteristics while de-emphasizing others. These critique sessions also seem to attune students to issues of personal choice and aesthetics. For example, one-sixth-grade student related how she became "addicted to frequency" after initially puzzling about how other students had created displays that looked very different from the case-value display that her team had produced (Shinohara and Lehrer 2017). Teachers support these critiques by encouraging students to take on perspectives other than their own, by asking students to trace a set of values from one display into another to emphasize how the shape of the subset changes with shifts in the conceptual tools employed to visualize them, and by posing questions or highlighting student contributions about what would happen to the shape of the data under different imagined scenarios (e.g., if more measurements were closer to the center value, if there were more extreme values). Teachers ask students to relate the shapes of the data displays to the data-generating process: Why are some displays (nearly) symmetric? Why is there a central tendency? What could be producing different values of measure, given that the object being measured is a constant length? What could happen if the process were repeated (for example, if the same object were measured by other students in the school)?

In summary, students are introduced to visualizing distribution by inventing representations of the outcomes of a firsthand, repeated process: measuring the same attribute of an object. Invented representations are compared to discern how the mathematical concepts of order, frequency, interval, and measurement scale result in different shapes for the same collection of values. As instruction proceeds, students' ways of generating visualizations of data stabilize. Nonetheless, as with statisticians, there are always elements of personal preference when students decide what to highlight and background in a data visualization.

### 3 Statistics as Measures of Distribution

Students are introduced to statistics as measures of distribution by again inventing and critiquing, albeit with the emphasis now on developing methods to arrive at estimates of the true length (or other attribute measured) and of the tendency of the measured values to agree (variability). The former corresponds to statistics of center and the latter to statistics of variance. During critique, students who did not create the method try to employ it to obtain a measure, with an eye again on what the invented methods attend to about the collection of values. Students typically rely on visualizations of the data that either they or others have previously developed to guide their construction of measures (Lehrer et al. 2011).

Invented measures of center usually include some of the canonical approaches, such as the most frequently occurring value (mode) or the middle-ranked value (median). The mean is not usually generated unless students have already learned about it. Other, less conventional solutions, such as the mid-range, are sometimes proposed. Student-invented methods for measure of variability also tend to bear some relation to canonical measures. For instance, some student methods emphasize deviation from a central value, such as the sample median. Others rely on counts of cases in regions that define a center clump of data, rather like a conventional interquartile range, or on the difference between extreme case values of the distribution, a kind of range statistic. Most student solutions bear some thematic relation to these classes of conventional solution. Unsurprisingly, no student has ever created an area model, as in the conventional definition of sample variance with squared deviations.

During class critique of these inventions, values of clarity and robustness often emerge. Students who are not authors struggle to follow the methods invented by peers and to imagine transformations of the distribution that will prove problematic for the invented measure (Lehrer and Kim 2009). For example, one student solution employed the sum of deviations from the median, but this measure did not result in values that reflected the changes in the variability that students observed when they switched from using the 15-cm ruler to the meter stick. The authors of the measure responded to this critique by revising their method so that it now employed the absolute value of deviations. This new solution reflected their intuition that the direction of difference from the median did not really matter. After this modification, a teacher asked what would happen if there were an increase in the sample size (e.g., if there were more measures) taken only with the meter stick. (The number of measures with the ruler would remain the same.) The invented statistic of variability again failed to conform to students' expectation that there would be very different variabilities in the distributions of values for the meter stick vs the ruler. Eventually, the class settled on a "per" solution that involved dividing the sum of deviations by the number of measures. Now the values of the invented statistic once again corresponded to the state of variability that the displays seemed to communicate.

Other student critiques that we commonly hear are sensitive to the context of repeated measures. For instance, many students criticize the range statistic for its reliance on the "two worst measurers." Teachers again play an instrumental role by inviting students to consider issues such as: What it is about a distribution that is indicated by a statistic, whether values of a statistic change in sensible ways with observed



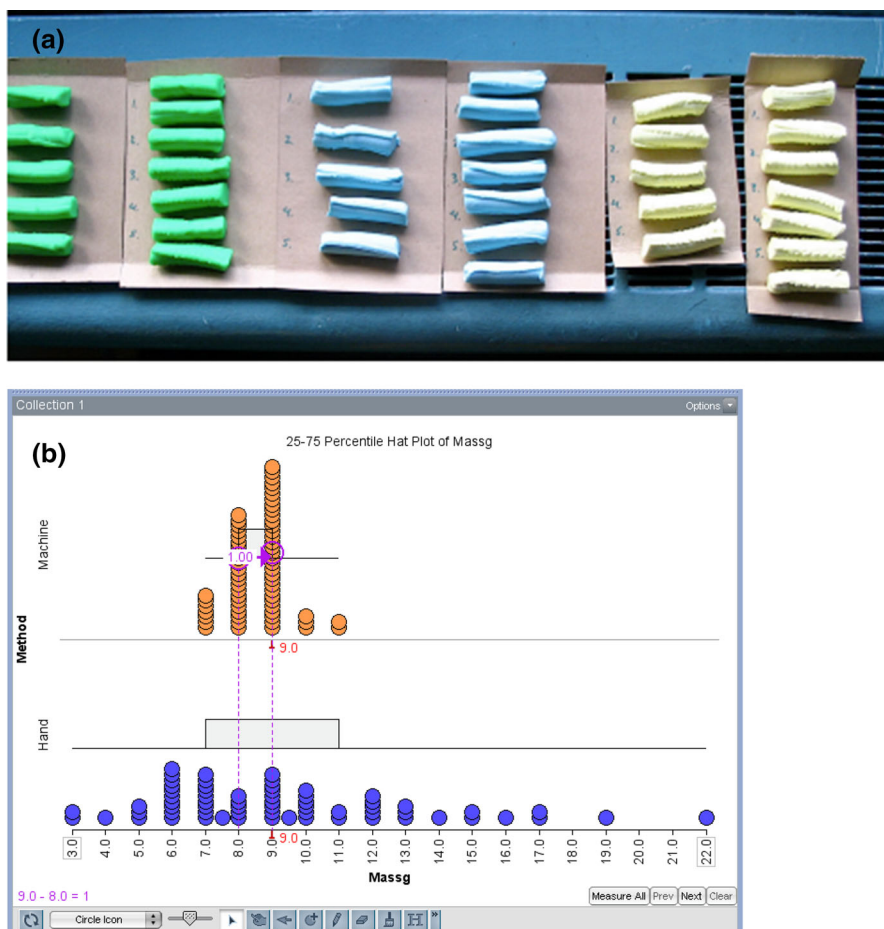
changes in the shape of a distribution, and whether the statistic is robust in the sense of generalizing to new imagined distributions.

Both invention and critique are critical to inducting students into the mathematical practice of measuring characteristics of distributions. Although statistics of central tendency and variability are usually taught directly, students often fail to understand the foundation for these measures. Accordingly, statistics come to be regarded as calculations to be performed routinely on batches of data. In contrast, the invention-and-critique routine invites students to grapple with the characteristics of distributions that are measured by statistics. Once students have done so, we introduce conventional measures, such as the mean, interquartile range, and average deviation. But students' experiences with trying to generate their own statistics help them regard these canonical measures as solutions to the problem of describing properties of distribution as quantities. These quantities are regarded as supplementing what is communicated by the visualizations of the distribution.

To continue the development of visualization and measure of distribution, we introduce a new signal and noise process, based on contexts of production (Konold and Harradine 2014). For example, students attempt to manufacture "candies" of a standard size out of clay. They employ different methods of fabrication (such as a machined form for extrusion vs hand crafting) (Fig. 2a). Or, students might use different methods for producing packages intended to hold a standard number of toothpicks. Manufacturing processes like these provide another example of a signal and noise process. Here, measures of central tendency are interpreted as target values of the process (e.g., number of toothpicks in each package), and measures of variability indicate the consistency of the products generated (e.g., variability in the number or mass of toothpicks across packages). These shifts in context for interpreting statistics help students understand that the same measures of distribution can be interpreted differently when the long-term generating process changes. Grasping this commonality is an example of mathematical generalization, in this case, extending the meaning of measures of distribution.

Students use a digital tool, TinkerPlots (Konold 2007; Konold and Miller 2011), to structure their data in ways that they have previously invented with paper and pen. The software tool provides greater flexibility and efficiency, now that students have had the opportunity to recognize some of the challenges of visualization that TinkerPlots resolves. For example, the tool has a drag operation that renders the batch of measurements continuously, so that gaps in the data are immediately visible. Recall that the value of this form of rendering is not always appreciated by students in their initial inventions of data visualizations. TinkerPlots can also be employed to generate summary statistics of central tendency and variability. Now these statistics are interpreted as measures of the target value of the process (e.g., the target value of each package) and of the consistency of product (how much does the mass of each box vary?). Figure 2b displays one of these comparisons between methods of production.

Teachers often initiate this context by introducing commercial standards that govern the production of consumables. For instance, state regulations require that the average mass of a sample of products is the target mass displayed on each package. No single package may deviate by more than five percent from this target value. The



**Fig. 2** a Production of Playdough “Fruit Bars,” b TinkerPlots visualization of different methods of production (color figure online)

notion of quality control further contextualizes practices of visualizing and measuring distribution.

## 4 Investigating Chance

Next teachers introduce the role of chance in signal and noise processes. These are not easy ideas for students to grasp. Their initial reasoning about probability is often based on single outcomes rather than repetition of a long-term process. Students also typically believe that random implies a complete absence of structure, rather than an inability to predict the outcome of any particular trial (Metz 1998). When they are deciding whether it is legitimate to accumulate outcomes, students often accept trials that are governed by different probabilities as equivalent, reasoning that because the

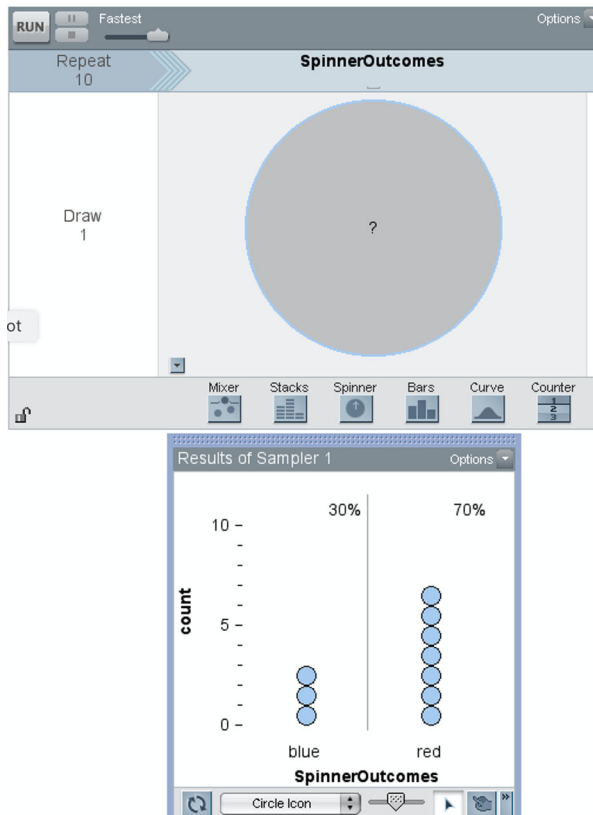
outcomes in each trial are “random,” they can be treated as sufficiently alike (Horvath and Lehrer 1998). For example, some students will happily accumulate the outcome of a vigorous throw of a die with one in which the thrower attempts and occasionally succeeds in adjusting the throw to achieve a predetermined outcome. They argue that this is acceptable because the biased throw is not always successful. These prior conceptions and others like them suggest that students also need experiences with chance and probability that involve stochastic processes.

Therefore, we involve students in investigating the behavior of simple random devices to develop stronger intuitions about the meanings of chance as a prelude to introducing chance as a model of the variability they have observed in the measure and production contexts. Teachers initiate this process with simple, hand-held, two-color (e.g., red, green) equal-area spinners. We ask each student to predict the result of applying hand force to the needle of a spinner ten times. Students also are asked to predict whether the direction of spin (clockwise or counterclockwise) will matter.

Typical student predictions include that exactly five of each color will result, that nine of their favorite color will result, that direction of spin does matter, and that the starting color will influence the final outcome. But as students experiment with spinners, they notice that not all of their predictions are sustained. We aggregate batches of ten spins, thus growing the sample, and ask students what they notice (Horvath and Lehrer 1998). The aim is to help students see that although they cannot predict any particular outcome with confidence, the longer-term pattern of outcomes is predictable and reflects the structure of the spinner. Noticing that changing the direction of spin does not affect the pattern of outcomes makes the concept of trial more visible and invites more careful consideration of what constitutes a repetition of a chance process.

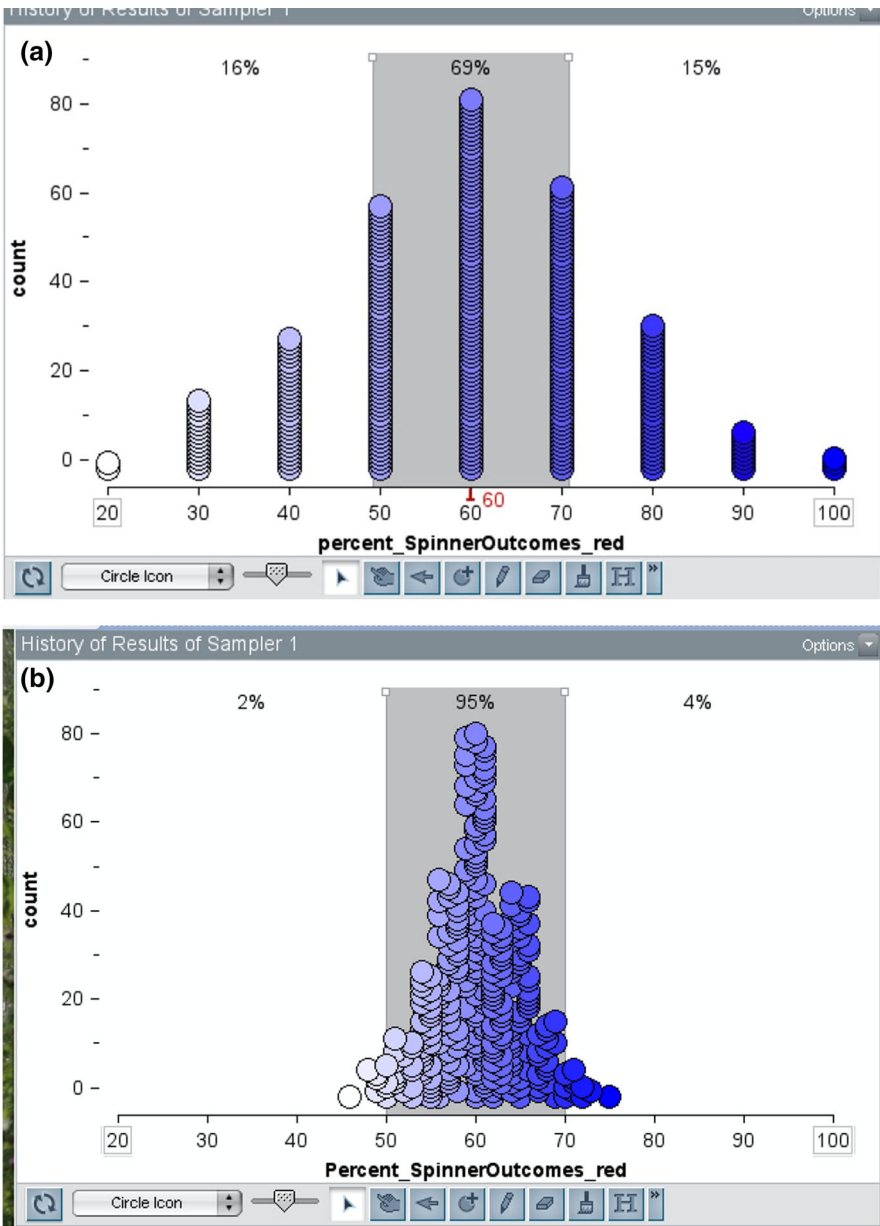
We do not recommend shortchanging the work with physical spinners in favor of computational simulations. Many of children’s intuitions about chance are rooted in ideas about their own effort, so it is important to provide time for some of these ideas to lose their persuasiveness. But once students have conducted a series of investigations with spinners, we introduce TinkerPlots’ random generators, which include a spinner representation, as analogs of the processes that students have just experienced physically. We explain to students that these random generators work like their spinners, but are faster and include built-in records of outcomes. We again explore two-region spinners but now encourage students to design regions with different areas. Students design “mystery” spinners with invisible partitions, and a partner tries to infer the structure of the spinner from repetitions of its behavior. Students find counterparts of their previous experiences with signal and noise in the structure of the spinner (signal) and chance departures (noise). Figure 3 shows a mystery spinner and a display of the outcomes of ten repetitions of its spins.

As students become familiar with the operation of spinners, we introduce the idea of using a statistic, such as percent red, to summarize the outcomes of a sample, as indicated in Fig. 3. Doing so sets the stage for investigating the properties of an empirical sampling distribution. For instance, students investigate the distribution of sample statistics over repeated experiments (e.g., 300 repetitions) and with different sample sizes (e.g., 10, 100). A result of one such investigation for a 60–40 spinner is displayed in Fig. 4a, b. Students redeploy their knowledge of statistics to indicate characteristics of the sampling distribution. They compare the central tendency and



**Fig. 3** Mystery spinner and outcomes of ten repetitions (color figure online)

variability of the two sampling distributions by finding the percentage of samples that lies between two values of the statistic, percent red. Teachers again play a key supporting role, asking students to relate a sample statistic to particular outcomes in a sample (e.g., 60% red means that six of the ten outcomes were red), to consider the meaning of particular cases (e.g., a dot in the display represents the percent red obtained in one sample of size 10 or 100), and to explain why the center remains invariant but the variability changes. Students typically relate the invariance of center to the unchanging structure of the spinner—a signal. But accounting for shifts in variability is more challenging. For example, one student explained that the chances of a statistic of 20% red for a 60% red spinner was low for a small sample (10 spins), but if one thought about it for a larger sample (size 100), then this low probability would now have to be repeated an additional nine times to obtain a sample of size 100 consisting of 20% red outcomes. Although this explanation parses trials that constitute the sample of 100 in ways that contravene conventional ways of thinking about this probability, nonetheless it does provide purchase for the student on a challenging idea (Lehrer 2017).



**Fig. 4 a** Empirical sampling distribution from 300 samples with sample size 10, **b** empirical sampling distribution from 300 samples with sample size 100 (color figure online)

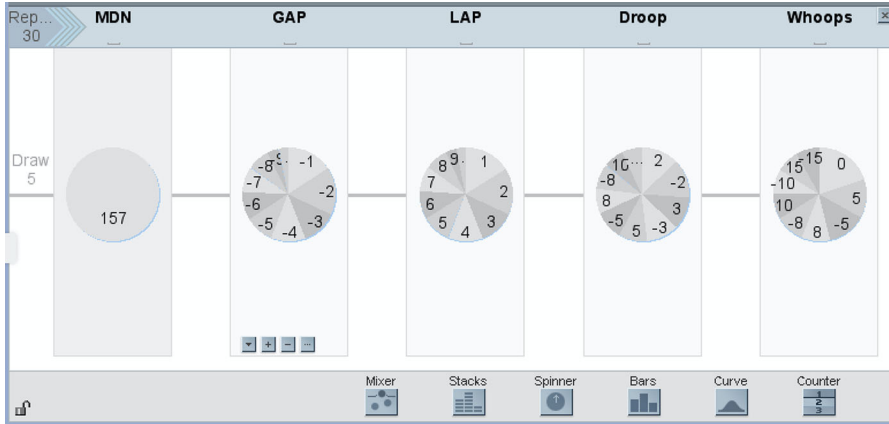
In sum, as students investigate the behavior of chance devices, they begin to develop an appreciation of chance that is based on repetition of a process and probability as an estimate of the signal produced by a chance process. They explain effects of sample

size on statistics of center and variability of a sampling distribution as reflecting the operation of signal, which tends to produce an invariant center, and change in variability due to the role of larger sample sizes in reducing noise around the signal. As they think about samples nested within sampling distributions, they begin to recast single samples as outcomes of a particular segment of an ongoing process—and simultaneously, as one of many such segments. This image of sample is further developed and stabilized as students begin to model variability.

## 5 Modeling Variability

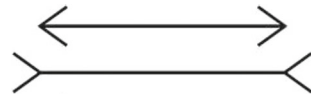
After students have received extended experience with more routine practices of visualizing and measuring variability of sample and sampling distributions, including those generated by random devices, we introduce them to modeling variability. As in earlier activities, students innovate by inventing and revising models and participate in critique by judging the rationale and behavior of the models invented by other students. Students re-envision the variability in repeated measure and production processes they have experienced firsthand as being due to the composition of fixed signal and random error. For example, Fig. 5 shows a student model that accounts for the variability in their class sample of repeated measures of their teacher's arm-span. The outcome of the first device is the sample median, intended to estimate (the "best guess of") the true value of her arm-span. The second random device designates magnitudes and probabilities (areas of the spinner) of "gaps" that occurred when the student iterated the ruler. Smaller magnitudes of gaps are judged as more likely than larger magnitudes. The magnitudes are assigned a negative value because gaps are unmeasured space that result in underestimates of the measure. The third random device designates magnitudes and probabilities of "overlaps" that occur when the endpoint of one iteration of the ruler overlaps with the starting point of the next iteration. The magnitudes are assigned a positive value reflecting that this form of error creates over-estimates of the measure, because the same space is counted more than once. The fourth random device represents both under- and over-estimates that result from the teacher becoming fatigued and repositioning her outstretched arms, an attribute that the students termed "slack." The last device represents probabilities and magnitudes of mis-calculations. The outcomes from each device are added to generate a simulated measurement value, and the process is repeated 30 times to represent a classroom of student-measurers and other participants (a few adult volunteers).

Following these and other challenges in model invention and revision, students participate in a psychophysics experiment. They mark the midpoint of the equal-length line segments portrayed in Fig. 6. The arrows at the ends of the line segments reliably bias the judgment of midpoint, and the challenge is to determine whether or not students in the class can overcome the effects of the (Mueller) illusion. The sixth graders, of course, are usually convinced that they can do so. To establish a baseline for inference, students first model the individual differences in an unbiased condition (the same length line segments are presented without any arrows), and in light of this model, make an inference about whether or not the class, on average, has indeed overcome the illusion.



**Fig. 5** Student model of fixed and random components of repeated measure of their teacher’s arm-span (color figure online)

**Fig. 6** Equal-length segments and the Mueller illusion



The Mueller illusion establishes a transition to natural variation. This shift is next continued by asking students to consider data from a plant laboratory where there was a concern about whether the day of planting was influencing the growth of batches of plants. Given standardized conditions of growth in the laboratory, the day of growth should not matter, and indications that growth was dependent on day of planting would necessitate careful scrutiny of standard conditions. Students are enlisted as “consultants” to the laboratory director, and they develop models that simulate outcomes that can be attributed to chance to make an inference about whether or not the day of planting matters. Modeling and inference now require envisioning plant growth as a repeated process, albeit in the case of natural variation, the process is governed by variables that are hidden from students’ view.

The teacher again plays a key role by asking students to explain how their beliefs about the process are reflected in their model, to use their model to trace how one simulated value is obtained, to consider what will happen if they run their model again, and to decide how they know whether their model is “good” (Wisittanawat and Lehrer 2018). During their participation in this sequence of model eliciting activities, students’ judgments about the quality of their models increasingly incorporate model-based sampling distributions of model parameters, such as the median of the simulated sample generated by the model and the interquartile range as an indicator of sample variability. Models considered good are those in which the values of the statistics of the empirical sample fit within the centers of the sampling distribution of the statistics generated by the model. For instance, the value of the IQR in the class sample is within the center clump of the simulated sample IQRs generated by the model.

To illustrate how engaging in modeling influences students’ perspectives on sample and inference in light of uncertainty (Ben-Zvi et al. 2012; Makar and Rubin 2017),

we next present episodes from a public school sixth-grade classroom (the families of most students qualified for lower cost or free lunch).

### *Constructing Models*

As noted, model construction was supported by students' participation in variability-generating processes, partly because processes involving signal and noise were intelligible to students and thus supported their reasoning about correspondences between the structure of a model and the process. For example, two students measured the perimeter of a table and discussed potential resulting sources of error. They agreed that ruler iteration error, "gaps and laps," should be included in their model, and then went on to consider other sources of error (student names are pseudonyms).

Brianna: And then we also have, like, calculation errors and false starts.

Cameron: How would we graph—, I mean, what is a false start, anyway?

Brianna: Like you have the ruler, but you start at the ruler edge, but the ruler might be a little bit after it, so you get, like, half a centimeter off.

Cameron: So, then it would not be 33, it'd be 16.5, because it'd be half a centimeter off?

Brianna: Yeah, it might be a whole one, because on the ruler that we had, there was half a centimeter on one side, and half a centimeter on the other side, so it might be 33 still, and I think we subtract 33.

Cameron: Yeah, because if you get a false start, you're gonna miss.

This example illustrates how making sense of sources of error (e.g., what is a false start?) with which participants had firsthand experience assisted generation of components of the model, typically resulting in variations of models like the one depicted in Fig. 5.

### *Possible Values and the Development of a Hierarchical Image of Sample*

One of the affordances of models like those depicted in Fig. 5 is that with repeated runs of the model, students observed that the simulation sometimes generated outcomes that were consistent with the variability-generating process, but were not case values in the empirical sample. For example, we next see students contesting the design a model of a fictitious production process that involved outputs of cookies of varying diameters. The argument concerned whether the model could include an output, a cookie diameter of 9 cm, that was not represented in the empirical sample of cookies.

Students: Take away – 1 in the spinner

Teacher: Why?

Joash: Because there's no 9 in this [the original sample].

Garth: Yeah, but that doesn't mean 9 is impossible.

Several students urged the inventors of the model to revise it so that magnitudes and probabilities of overshooting or undershooting the targeted diameter (the errors) would generate only simulated values that also appeared in the empirical sample. Garth (an inventor of the model under scrutiny) and a few other students argued that the simulated value under contest was plausible—it *could* have been produced, but by chance, it wasn't. The model "focused on the probability of messing up," so to the extent to which error magnitudes and probabilities were plausible, one would have to accept simulated values generated by the model as possible values.



“Possible values” eventually assumed increasing prominence with students, to a point where they began to consider empirical samples as simultaneously a collection of outcomes observed in the world and a member of a potentially infinite collection of samples. We see another discussion of this relationship as students modeled the first, “no illusion” condition of the eye illusion experiment. The students were trying to account for individual difference variability and a remarkably accurate central tendency (a median of 100u estimated as the midpoint of a 200u line). During critique, the class considered a student invention that treated the individual differences as noise. Once again, this invented model-generated values that were not present in the empirical sample. Seizing on this, the teacher highlighted the generation of one such value (95) and asked whether it was a flaw in the model that the value did not exist in the sample. In response, one of the model inventors appealed to the plausibility of the value as a possible value (of the random process), a position that was ratified by many of his classmates.

Teacher: Okay, but there is a gap at 95, and they have a  $-5$ , which means they can get—

James: Because it’s possible. There’s a chance.

Teacher: But it didn’t happen when we really did it, so it shouldn’t be up there, should it?

Students: It should, because it’s still possible.

The teacher playfully continued to push her objection in an attempt to elicit further clarification:

Teacher: Let’s vote. Who thinks they should not have  $-5$  up there? Who’s with me? [A few students raised their hands.] Because it didn’t happen?

Student: So?

Abe: It’s possible.

Savannah: It’s still possible.

Student: It’s plausible.

Student: Just because it didn’t happen—

Teacher: Based on the data, it’s impossible. It didn’t happen [pointing at the empty spot at 95 in the observed sample].

Savannah: Well, it’s between the two numbers that did happen, so it’s possible.

At this point in the conversation, a student framed the empirical sample as just one realization of the variability generating process instantiated by the model:

Brianna: That’s just one data set.

Teacher: Oh, so Brianna said this is just one data set.

Students: Yes.

Teacher: So, Brianna, why do I care that that’s just one data set? What does that tell me?

Brianna: Because not everyone is gonna get every single number. If we run it more times, there’re more possibilities.

Teacher: What do you mean run it more times? Run the model?

Brianna: Well, if we were to do the line more times, again and again, if Mr. B’s class does it, then it will be—

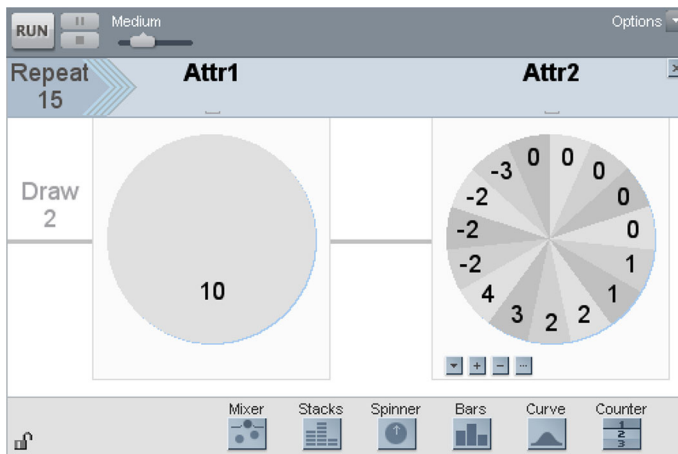


Fig. 7 A resampling model of cookie production (color figure online)

James: Like if we do the line again.

Teacher: Oh, you mean if I collect the real data again, it's possible to get 95, so since it's possible for it to really happen in real life, my model might want to represent that happening?

Students: Yes.

#### *Alternative Perspectives: Resampling Models*

Although the majority of student pairs constructed signal–noise models through analysis of sources of variability, some teams adopted an alternative perspective of resampling the empirical sample to account for the outcomes of a production, as suggested earlier by their objections to models that generated values not included in the empirical sample. For example, Fig. 7 displays the model constructed by a pair who generated signal and random error in a way that reflected the proportion of outcomes in a fictitious empirical sample of cookies.

This model provoked immediate contest, with some students objecting on the basis of the model's exclusive reliance on the empirical sample, which ruled out other possible outcomes consistent with the process. For example,

Joash: They copied down this [the original data] ... which makes you wonder how did they get this data.

Jean: That's one batch of cookies. You can get 9 at any other time, because your scoop is not mathematically [perfect].

After further discussion of this perspective, the teacher continued:

Teacher: Okay, so, what is good about this, though? Abe?

Abe: Well, what's good about it is that it expresses the true probability that it can get whatever numbers that are on there [the original sample].

.... [continued conversation]

Teacher: Well, can I ask what the true probability of getting 10 is, Abe?

Abe: Well there's five 0's, so it should be 33%.

Students:  $1/3$ .

.... [continued conversation]

Teacher: So, my question is, is every time I make a batch, I'm gonna have exactly five out fifteen 10's?

Students: No.

Garth: Not every time, that could happen.

Teacher: Okay.

The teacher established that the students did not expect the model to copy, but rather to approximate the empirical sample, even though it was guided exclusively by the estimated probabilities of outcomes in the sample. We speculate that the appeal of this model was based in part on the fictitious nature of the sample, so even though sources of error were identified in the problem scenario, students did not have firsthand experience of them. The resampling perspective on modeling gained traction as the class confronted problems of natural variation. For example, a student creating a resampling model of the estimates of midpoint of the line in the unbiased condition of the eye illusion experiment clarified that ignorance of the process motivated consideration of resampling: "We kinda did like what they did yesterday, because there's not really a process for doing this, so we just sampled it, and we did the values of all the errors of the original" (an analog to the resampling model of the production process invented by peers). Knowledge limitations about process were increasingly prominent in the challenge of modeling the difference between the number of leaves after 18 days of growth for two batches of plants, one planted earlier than the other. Now students typically resorted to putting all cases into a common "mixer" and sampling with random replacement to arbitrarily assign a case value to one time of planting or to the other. The following episode exemplifies some of the conversation about students' rationale:

Teacher: So, is there chance involved in whether our plant grows 5 leaves or 8 leaves by the 18th day?

Students: Yes.

Teacher: So, we are assuming that our original data set, there're chance involved in that, right? So, we're continuing to assume that there's some chance involved, because we're using a chance device.

Student: Right.

Teacher: And, we're using the original data set, we poured it all into the mixer. Why?

Joash: So we can see what happens at different times with the same numbers.

Teacher: Shreya, why do we pour all that information back into the mixer, and not create a spinner with errors?

Shreya: That's all we know.

Teacher: That's all we know, right?

Joash: We don't know the errors, if there are any.

Teacher: Yeah, and what we do know is really complicated, and so we don't really want to go through the process of quantifying it.

Resampling was not met with unanimous acceptance. Drawing upon their experiences with constructing sampling distributions, some students pointed out that the samples were relatively small:

Cameron: Um, like, you have to do this again. You can't, it's like with the boy and the girl thing [reference to a problem posed using birth data from two counties in North Carolina]. If you have one kid born, and it's a boy, it's gonna be 100%, but then if you do it like 100 times, and then if you have 100 births ... then you have a [gesturing], but if you only do the experiment once and then you put all that data into a mixer, then your probability, your probability of getting something is kind of—

Teacher: So, are you saying if we had done more plants, we'd grown more plants, instead of just 36, we would have better numbers?

Cameron: Do the experiment again or something.

Teacher: Do another experiment. Do the same experiment multiple times.

Students went on to consider some of the reasons the scientists had not (yet) repeated the experiment but they eventually settled on employing the data that were available to construct an empirical sampling distribution of mean differences to infer whether the day of planting did make a difference in this measure of plant growth. We next illustrate how students employed models and model-generated simulation of sampling distributions to guide inference.

### *Model-Based Inference*

After students decided whether or not a model was good, generally through recourse to the behavior of the model-generated sampling distribution of statistics of central tendency and of variability, such as the IQR, the teacher and researcher (RL) asked students to consider the likelihood of values of these statistics at or in excess of some threshold value. The aim was to help students establish routines for considering the probability of a value of a statistic by locating it within the model's simulated sampling distribution of that statistic, as illustrated in the following exchange.

RL: So how likely is it that the median, just by chance, would be more than 510?

Savannah: Still a little bit.

Joash: Out of this?

RL: Ah ha.

John: Turn on the percent and I'll tell you.

RL: You want to guess?

Teacher: Let's guess before we turn it on.

.... [continued conversation as students estimate]

Teacher: Write down your vote in your journal. ... What's the chance that it lands, that a class, that we get a median of higher than 510. 7%? 8?

Joash: Yeah it is 5%. I found out that it's 5%.

Renee: Yeah, I did the math. It's 5%

To initiate the logic of inference, the teacher went on to pose a scenario in which another class claimed that they measured the same table and the median of their sample of measurements was an extreme value in the sampling distribution generated by this class's model.

Teacher: Do you think Mr. B's class got mixed up and measured a different table, or do you think that just by chance, they could have gotten 517?

Joash: On this data [looking at sampling distribution], it looks like they messed up... So how will the median be 517 if that will be an outlier [in the sampling distribution]?

Other students imagined a shift or translation of the distribution of measurements in the second class, further buttressing their rejection of the claim that both classes had measured the same table's perimeter.

Joash: If this is the correct model—

Teacher: We're assuming it is, yeah.

James B: On there, our big pile is around like 503, 501, but theirs would be like their middle part would be around 517.

Teacher: Oh so not—

Joash: Theirs shifted [gesturing shifting to the right].

Garth: Their whole data just shifted.

As students continued to construct models of different scenarios, the teacher cultivated reasoning about the sampling distribution of statistics of simulated samples generated by models as a way of deploying chance to inform inference. Consequently, this approach gradually achieved more prominence, as illustrated in the following episode. Here, a teacher invited students to consider sample values in the bias condition that they would take as evidence of overcoming the Muller illusion. In this episode, the students appeal to any value of the median within the simulated values of the median generated by their model of unbiased judgment.

Teacher: Okay, so Lily, if Mr. B's class says we've got a median of 102, do you think they overcame the illusion?

Lani: Yes.

Teacher: Yeah? 101?

Lani: Yeah.

Teacher: 103?

Lani: Yes.

Renee: Pushing it, but yeah.

Teacher: You think they overcame the illusion? 97?

Lani: Yeah.

Attempting to shift students' focus from inclusion or exclusion in the distribution to probability, the teacher then challenged students to think of the probability of a value that was in the range generated by the simulated sampling distribution of the medians

Teacher: So, in your opinion, if they get any of those numbers that you got, they overcame the illusion. Okay. Even this one that only happens, I just want to see, 97 happens less than a percent of the time? If they got 97, you think they overcame the illusion or they fell for the illusion? Did it shift it?

Lani: They're probably affected.

Teacher: Ha, I don't know. Talk about it. What would make you confident? What percent would have to be there for you to feel confident that they overcame the illusion?

The teacher's efforts to consider a statistic in light of its probability in a sampling distribution began to be appropriated by students. They continued to model natural

variation to determine whether or not the difference between sample means observed in number of leaves of plants in batches 18 days after initial planting, but at two different times, was likely merely a manifestation of chance.

Teacher: What was our original data set's difference?

Students: .72

Teacher: .72 leaves. Do we think that happened by chance or do you think the original assumption, day doesn't matter is [wrong]?

... [continued conversation]

Teacher: What does it mean for it to have only happened 4% of the time by chance? What does that mean?

Renee: Probably not a chance happening [the difference is due to day of planting].

Savannah: It's a very low chance.

Reflecting concerns raised earlier during the discussion about sample size, several students suggested that perhaps the low probability of a difference as large or larger than that observed might be erased if the samples were larger:

Brianna: If I were telling the scientist what to do, then I would tell them to collect more data, because it's still possible to get 0.72, but it's highly unlikely, and I think would be more helpful to collect more data.

Joash: It'd get knocked out because of the larger sample size.

Brianna: Yeah.

Renee put it more gently, but she also anchored the model to what she understood about the situation, questioning the inference based on the model:

Renee: I'm thinking, well, the data, it looks like day should matter, because it's such a small percentage of it happening by chance, like there's 4% chance that it happened by chance. But the day really shouldn't matter if all the conditions are the same and they are growing inside, the same temperature, humidity, and everything, but I want to see like three more testings or something. Just a little bit more.

Across these episodes of classroom interaction, it appears evident that engaging students in modeling tangible processes involving signal and noise provided a productive entrée to conceiving of variability as generated by a long-term process with fixed and random components. The insights that students generated through firsthand experience of signal and noise, in contexts of repeated measure and manufacture, helped them navigate toward new understandings of sample and sampling distributions and of inference in light of uncertainty (Manor Braham and Ben-Zvi 2015). The bridge between these statistical concepts and students' experiences was built by representing these experiences as models that simulated the variability of outcomes observed in the world as compositions of signal and noise. Modeling processes with "possible values" anchored images of sample as simultaneously representing both one set of observations of the outcomes of a process and as one of a potentially infinite series of such samples (Lehrer 2017). Students' invention of models involving resampling as a stand-in for unobservable process involved in natural variability helped bridge direct, firsthand experience of variability with variability in the natural world.

## 6 Discussion

Practices like modeling are the generative means by which disciplines create and revise knowledge. We have described one pathway through which comparatively young students participated in the kinds of practices that allow professionals to get a grip on variability. Much of the process of data modeling (Lehrer and Romberg 1996; Lehrer and English 2018) can be made accessible to students by engaging them in the invention, revision, and contest of ways to visualize, measure and model variability. Through this work they can come to grasp some of the central concepts and forms of activity that generations of scientists and statisticians have developed. We described a progression of variability-generating processes that helps students form images of variability as emerging from a long-term, stochastic process. Students come to regard visualizations of variability as being determined by choices made by designers about count, order, interval, and scale. Increasingly, students conceive of variability as measured by attending to particular characteristics of the cases that comprise a collection, modeled by compositions of fixed and random components.

This modeling sequence provided students with some useful tools for pursuing understanding of the natural world. For instance, sixth-grade students subsequently employed these ideas to pursue questions about the ecology of a local pond. They sought to estimate the extent to which the relative abundance of target organisms (e.g., caddisfly, dragonfly nymphs) could be considered as arising merely from chance variation. They wondered: Were apparently higher counts of the organism found in the same volume of pond water due to the location of the sample, or were they perhaps due only to the random variation one would expect from sample to sample? Confronted with warranting claims about the relative abundance of organisms in particular partitions of the pond, students decided to pool all their counts of the organism they had observed throughout the pond in a random mixer. Then, they repeatedly drew a sample at random (with replacement) and created an empirical sampling distribution. Students then made judgments about new samples drawn from different locations in relation to this empirical distribution. Although this approach has its limits, it nonetheless signals that students were seeking to extend their experiences of modeling to new realms of experience. We do not argue that the sequence delivers a complete toolkit for modeling variability, but it does afford young students with a basic conceptual repertoire that can support them in pursuing their own questions, rather than being constrained to those dictated by curricula or educators.

The sixth-grade teacher's support of student invention and revision of models, and her efforts to help students deploy modeling to support inference, exemplify some of the instructional practices through which teachers bring data modeling to life in classrooms. As teachers accumulate experience, they also contribute to curricular innovation by contributing to a "Teachers Corner" that features additions and amendments to the baseline curriculum. Some of these contributions include:

- guidelines for articulating gesture to highlight mathematically relevant characteristics of data displays for ESL students,
- organizers for expanding the instructional reach of formative assessments by attending to typical patterns in students' ways of thinking,

- using TinkerPlots to help students think about the mean as a fulcrum of deviations of case values from it,
- contrasts between measurement precision and accuracy (i.e., bias),
- guiding exploration and discussion of students' beliefs about personal agency and randomness,
- imports of sports data (e.g., fantasy football) to extend ideas of statistics as measures, and
- extensions to new production processes, such as making candies.

Teacher innovation also includes authoring for professional publication. For instance, some schools expanded the cycle of innovation and critique of data displays in the first phase of instruction to engage in school-wide norm-setting about how to conduct mathematically productive conversation. Teachers coauthored an article about their experiences for a wider professional audience (Tapee et al. 2019). The lead authors were teachers we never interacted with, evidence of the kind of appropriation that indicates a robust design. These innovations and extensions are but a small sample, but they illustrate how our initial design has been elaborated and revised in response to teacher feedback and improvement. As the examples suggest, the design is not fixed but is continually shaped by participants, especially as we learn more about students' propensities and characteristic ways of thinking, and as the social worlds of students and teachers change. When we first initiated a variant of this design, sport statistics were not easily accessed; now it is difficult to avoid being enveloped by them. Similarly, data generated by students could not be readily shared outside of the confines of school, and now large-scale data, including those of scientists, are more readily available. Both of these changes have generated more points of contact between student-generated, small-scale data and so-called big data. These changes in access and availability suggest augmented possibilities for data modeling. Nonetheless, we conjecture that the firsthand experiences of variability and chance that we have described will continue to be critical and will provide an important foundation to children's subsequent work with data farther from their own personal experience.

**Acknowledgements** Funding was provided by Australian Research Council (Grant No. DP180102333) and Institute of Education Sciences (Grant No. R305A110685).

## References

- Bakker A (2018) Design research in education. Routledge, New York
- Ben-Zvi D, Aridor K, Makar K, Bakker A (2012) Students' emergent articulation of uncertainty while making informal statistical inference. *ZDM* 44(7):913–925
- Ford MJ (2015) Educational implications of choosing “practice” to describe science in the next generation science standards. *Sci Educ* 99(6):1041–1048. <https://doi.org/10.1002/sce.21188>
- Ford MJ, Forman EA (2006) Redefining disciplinary learning in classroom contexts. *Rev Res Educ* 30:1–32
- Hestenes D (1982) Modeling games in the Newtonian world. *Am J Phys* 1 60(8):732–748
- Horvath JK, Lehrer R (1998) A model-based perspective on the development of children's understanding of chance and uncertainty. In: Lajoie SP (ed) *Reflections on statistics: learning, teaching, and assessment in grades K-12*. Routledge, New York, pp 121–148
- Jones RS, Lehrer R, Kim M-J (2017) Critiquing statistics in student and professional worlds. *Cogn Instr* 35(4):317–336. <https://doi.org/10.1080/07370008.2017.1358720>



- Konold C (1989) Informal conceptions of probability. *Cogn Instr* 6(1):59–98. [https://doi.org/10.1207/s1532690xci0601\\_3](https://doi.org/10.1207/s1532690xci0601_3)
- Konold C (2007) Designing a data tool for learners. In: Lovett MC, Shah P (eds) *Thinking with data*. Lawrence Erlbaum Associates, Taylor and Francis, New York, pp 267–292
- Konold C, Harradine A (2014) Contexts for highlighting signal and noise. In: Wassong T, Frischmeier D, Fischer PR, Hochmuth R, Bender P (eds) *Mit Werkzeugen Mathematik und Stochastik lernen: using tools for learning mathematics and statistics*. Springer, Wiesbaden, pp 237–250
- Konold C, Lehrer R (2008) Technology and mathematics education: an essay in honor of Jim Kaput. In: English LD (ed) *Handbook of international research in mathematics education*, 2nd edn. Taylor & Francis, Philadelphia, pp 49–72
- Konold C, Miller CD (2011) TinkerPlots. Dynamic data exploration. Key Curriculum Press, Emeryville
- Konold C, Higgins T, Russell SJ, Khalil K (2015) Data seen through different lenses. *Educ Stud Math* 88(3):305–325. <https://doi.org/10.1007/s10649-013-9529-8>
- Lehrer R (2017) Modeling signal-noise processes supports student construction of a hierarchical image of sample. *Stat Educ Res J* 16(2):64–85
- Lehrer R, English LD (2018) Introducing children to modeling variability. In: Ben-Zvi D, Makar K, Garfield J (eds) *International handbook of research in statistics education*. Springer International Publishing, Cham, pp 229–260
- Lehrer R, Kim MJ (2009) Structuring variability by negotiating its measure. *Math Educ Res J* 21(2):116–133
- Lehrer R, Romberg T (1996) Exploring children's data modeling. *Cogn Instr* 14(1):69–108. [https://doi.org/10.1207/s1532690xci1401\\_3](https://doi.org/10.1207/s1532690xci1401_3)
- Lehrer R, Schauble L (2007) Contrasting emerging conceptions of distribution in contexts of error and natural variation. In: Lovett MC, Shah P (eds) *Thinking with data*. Lawrence Erlbaum Associates, Taylor and Francis, New York, pp 149–176
- Lehrer R, Schauble L (2017) Children's conceptions of sampling in local ecosystems investigations. *Sci Educ* 101(6):968–984. <https://doi.org/10.1002/sce.2129>
- Lehrer R, Schauble L (2019) Learning to play the modeling game. In: Upmeier zu Belzen A, Kruger D, van Driel J (eds) *Toward a competence-based view on models and modeling in science education*. Springer, Cham, pp 221–236
- Lehrer R, Kim M-J, Jones RS (2011) Developing conceptions of statistics by designing measures of distribution. *ZDM Math Educ* 43(5):723–736. <https://doi.org/10.1007/s11858-011-0347-0>
- Lehrer R, Kim M-J, Ayers E, Wilson M (2014) Toward establishing a learning progression to support the development of statistical reasoning. In: Confrey J, Maloney AP, Nyuyen KH (eds) *Learning over time: learning trajectories in mathematics education*. Information Age Publishers, Charlotte, pp 31–60
- Makar K, Rubin A (2017) Learning about statistical inference. In: Ben-Zvi D, Makar K, Garfield J (eds) *International handbook of research in statistics education*. Springer International Publishing, Cham, pp 261–294
- Manor Braham H, Ben-Zvi D (2015) Students' articulations of uncertainty in informally exploring sampling distribution. In: Zieffler A, Fry E (eds) *Reasoning about uncertainty: learning and teaching informal inferential reasoning*. Catalyst Press, Minneapolis, pp 57–94
- Metz KE (1998) Emergent understanding and attribution of randomness: comparative analysis of the reasoning of primary grade children and undergraduates. *Cogn Instr* 16(3):285–265. [https://doi.org/10.1207/s1532690xci1603\\_3](https://doi.org/10.1207/s1532690xci1603_3)
- National Research Council (2012) *A framework for K-12 science education: practices, crosscutting concepts, and core ideas*. The National Academy of the Sciences, Washington, DC
- Nersessian NJ (2008) *Creating scientific concepts*. The MIT Press, Cambridge
- Petrosino AJ, Lehrer R, Schauble L (2003) Structuring error and experimental variation as distribution in the fourth grade. *Math Think Learn* 5(2–3):131–156. <https://doi.org/10.1080/10986065.2003.9679997>
- Pfannkuch M, Wild CJ (2000) Statistical thinking and statistical practice: themes gleaned from professional statisticians. *Stat Sci* 15(2):132–152. <https://doi.org/10.2307/2676728>
- Saldanha LA, Thompson PW (2002) Conceptions of sample and their relationship to statistical inference. *Educ Stud Math* 51(3):257–270
- Saldanha LA, Thompson PW (2014) Conceptual issues in understanding the inner logic of statistical inference: insights from two teaching experiments. *J Math Behav* 35:1–30

- Shinohara M, Lehrer R (2017, July) Narrating lines of practice: students' views of their participation in statistical practice. Paper presented at the tenth international research forum on statistical reasoning, thinking, and literacy (SRTL 10). Rotorua, New Zealand
- Tapee M, Cartmell T, Guthrie T, Kent LB (2019) Stop the silence. How to create a strategically social classroom. *Math Teach Middle Sch* 24(4):210–216
- Thompson PW, Liu Y, Saldanha L (2007) Intricacies of statistical inference and teachers' understanding of them. In: Lovett MC, Shah P (eds) *Thinking with data*. Lawrence Erlbaum Associates, Taylor and Francis, New York, pp 207–231
- Windschitl M, Thompson J, Braaten M (2008) Beyond the scientific method: model-based inquiry as a new paradigm of preference for school science investigations. *Science Education* 92(5):941–967. <https://doi.org/10.1002/sce.20259>
- Wisittanawat P, Lehrer R (2018) Teacher assistance in modeling chance processes. In: Sorto MA (ed) *Looking back, looking forward*. (Proceedings of the 10th international conference on the teaching of statistics, Kyoto, Japan, July). Voorburg, The Netherlands: International Statistical Institute. Accessed on 26 Feb 2020 from [https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10\\_2A2.pdf](https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_2A2.pdf)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.