MODELING SIGNAL-NOISE PROCESSES SUPPORTS STUDENT CONSTRUCTION OF A HIERARCHICAL IMAGE OF SAMPLE

RICHARD LEHRER
Vanderbilt University
rich.lehrer@vanderbilt.edu

ABSTRACT

Grade 6 (modal age 11) students invented and revised models of the variability generated as each measured the perimeter of a table in their classroom. To construct models, students represented variability as a linear composite of true measure (signal) and multiple sources of random error. Students revised models by developing sampling distributions of model-generated statistics to judge model fit and validity. After instruction, interviews with 12 students were conducted to learn how they conceived of relations among chance, modeling, and inference. Most students' inferences were guided by a hierarchical image of sample, a perspective constituted through their understandings of modeling variability as signal and noise.

Keywords: Statistics education research; Model; Model-fit; Sampling distribution

1. INTRODUCTION

1.1. THE ROLE OF MODELING IN STATISTICAL INFERENCE

One of the prominent agendas of statistics education is to support students' understanding of how inference can be made in light of variability (Garfield & Ben-Zvi, 2008). Formal approaches to inference emphasize modeling the outcomes of random processes as functions that generate probability densities and that regulate inference. However, people's grasp of the logic of this approach typically develops over a considerable period of time (Pfannkuch & Wild, 2000). One challenge to accomplishing understanding is that the logic of hypothesis testing is counterfactual. Another is that sustaining this logic requires maintaining a hierarchical image of sample in which samples are composed of visible cases that vary, and in which samples, in turn, are simultaneously members of an imagined collection that varies (Ben-Zvi, Bakker & Makar, 2015; Saldanha & Thompson, 2002, 2014; Thompson, Liu & Saldanha, 2007). A third challenge is that inference relies on intimate knowledge of the problem context, which is required to ground sense-making (Makar, Bakker, & Ben-Zvi, 2011). In light of these significant challenges to the formal foundations of inference, it is worthwhile to begin supporting students' initiation to these ideas via informal inference, an approach that emphasizes going beyond particular samples to make generalizations that explicitly recognize the uncertainty of the generalization. Uncertainty is often signaled via linguistic hedges such as "may," or references to a neighborhood of values (Makar & Rubin, 2009; Pfannkuch, 2011;

Pratt & Ainley, 2008). The informal inference approach clarifies for students the foundational role of inference in statistics without insisting on formalisms that may be poorly understood and that take considerable time to develop.

This research explores a way to expand young students' (age 11) articulations of informal inference by engaging them in an approximation to the professional practice of participating in a "dialog between models and data" (Cobb & Moore, 1997, p. 810). As described momentarily, the conjecture that motivates this study is that student invention and revision of models of variability can be rooted in their participation in accessible processes involving signal and noise (Konold & Harradine, 2014; Konold & Pollatsek, 2002; Lehrer, Kim, & Schauble, 2007). It is expected that these experiences will provide pathways for developing a hierarchical image of sample that, in turn, can guide informal inference.

A hierarchical image of sample integrates sample variability and sampling variability (Rubin, Bruce & Tenney, 1990). For example, students who observe a sample statistic with a hierarchical image in mind will likely make inferences about that statistic based on their anticipations of its sampling variability. In formal approaches, this anticipation is based on probability densities and functions. In contrast, the approach taken here was to engage students in inventing chance models of the variability that resulted from a repeated measure process in which they were the agents. Ideally, repeated measures can be modeled as a composition of fixed signal, the true measure, and variation arising from random error. Such models are cognitive tools that could support student reasoning about commonalities, the same signal, and differences, random error, from (simulated) sample to sample. Modeling this composition of signal and noise may help students envision sampling variability as sensible and even inevitable. By revising models, students also have means to observe which correspondences between components of the model and sampling variability seem to matter. If students believe that their model is a valid representation of the process, then the empirical sample constituted by their actual measurements can be seen as a prospective member of the potentially infinite collection of samples generated by the model. Hence, modeling is a potential pathway for envisioning the hierarchy of variable cases within a sample and an imagined collection of samples that vary. If students also have a means to augment their vision, such as a sampling distribution that makes explicit how a statistic tends to vary from sample to sample, then they have the semiotic means to expand the reach of informal inference to include probabilistic judgment.

To explore the potential of this form of modeling for expanding students' conceptions of sample and consequently, of informal inference, we conducted an eight-week design study in a sixth-grade classroom (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003). During the latter half of the design study, students invented and revised models, primarily those representing the variability of processes involving signal and noise. They used their models to make inferences about claims of change in process. At the completion of instruction, a sample of participating students responded to a three scenarios posed in a flexible interview. Student conceptions of, and the relations among, sample variability, sampling variability, and modeling were elicited, and we further attempted to learn how students understood the role of each in informal inference. Accordingly, research questions included:

> (1) How did students conceive of sample distribution and the role of sample size in estimating sample statistics?

(2) For a given sampling distribution, how did students conceive of cases, interpret statistics of center and variability, and predict and explain the influence of sample size on statistics of center and variability?

(3) Working with familiar contexts of signal and noise, how did students understand the operation of models? What criteria did they generate to decide about "good" models? Could they change model parameters to generate hypothetically different distributions?

(4) What role did models and sampling distributions of model statistics play in guiding students' inferences about new claims and associated statistics?

1.2 INSTRUCTION TO SUPPORT MODEL-BASED INFERENCE

The entire instructional sequence, including teacher-generated elaborations of lessons and supporting materials, is available (at modelingdata.org), so the description of the design and rationale of instruction is described here sparsely. The aim is to give the reader a sense of the curricular trajectory in which students participated; the research is directed at student reasoning post-instruction. However, student experiences were not simply a matter of posing curricular tasks. The classroom teacher had participated previously in related research aimed at describing and assessing student thinking about statistics and data during the late elementary/early middle school years (Lehrer, Kim, Ayers, & Wilson, 2014).

Making sense of sample variability The initial phase of instruction supported students in developing an image of a sample as arising from the repetition of a process (Saldanha & Thompson, 2002). Components of processes involving signal and noise are usually more visible to young students than those involving natural variation (Konold & Lehrer, 2008; Lehrer & Schauble, 2007). Every student used a 15-cm ruler to measure the perimeter of the classroom's "lost-and-found" table. Therefore, students had a hand in the data generating process, often an essential underpinning for interpreting data (McClain & Cobb, 2001; Lehrer & Schauble, 2002). Students speculated about sources of variability in their measures and considered what would happen if other classes also measured the same table with the same tool. The intention was to help students think of measurement as a process and to begin to identify plausible causes of variability, an important resource for learning to think statistically (Biehler, 1999; Prodromou & Pratt, 2006; Wild, 2006).

Working with the classroom batch of data, pairs of students used paper and markers to display the data in ways that would "help someone else see a trend or pattern that you noticed." Following this data display invention phase, whole-class critique focused on detecting what different displays "show and hide." The goal was to help students understand that the "shape of variability" arises from choices made by designers, and these choices are consequential for what one can readily "see" about variability (Lehrer et al., 2007; Lehrer & Schauble, 2004). Several student displays made visible a modal or center clump (Bakker, 2004), and the teacher highlighted these clumps by asking students to account for the bell-like shape. Here, again, the teacher sought to anchor variability to characteristics of process, including: signal, "what the perimeter of the table is," and noise, "different kinds of mistakes."

Students concluded their exploration of sample variability by inventing statistics to estimate both the "real perimeter" and "precision, or how much our measurements tended to agree." Collectively, students reviewed the inventions proposed by the class and considered

what the different statistics attended to about the sample. These conversations were important in fostering a view of a statistic as a measure, rather than merely as a computation (Lehrer & Kim, 2009; Lehrer, Kim, & Jones, 2011). Next, this cycle of constructing, visualizing, and measuring variability was extended to other accessible contexts that were characterized by visible sources of variability (noise) and stability (signal) and in which students could participate in firsthand generation of data. For example, students compared the consistency of products (e.g., "fruit sticks") they generated with different production processes—see modelingdata.org, Unit 4.

Extending variability to include chance Exploration of variability was next extended to include the role of chance. First, students investigated the behavior of simple chance devices with TinkerPlots (Konold, 2007; Konold & Miller, 2011). These investigations prompted students to consider relations between the structure of a chance device (such as the percent area of red in a two-color spinner) and the long-run relative frequencies or proportions of targeted events. A second investigation focused on sample-to-sample variability, initially noticed by students during their investigations of sample variability. Investigations included collecting 300 sample statistics for small (n=10 repetitions of the spinner), medium (n= 100), and large (n =1000) samples consisting of outcomes of 2-color spinners. Students explained what the median and IQR of the sampling distribution measured and discussed why the values of these statistics were either stable (median) or changing (IQR). These activities focused students on the relations among sample, sample statistic, sampling distribution, and statistics of sampling distributions.

Modeling variability of measurement After these excursions into chance and sampling, students again re-considered their measurement data, recalling their best estimate of the perimeter of the table and identifying sources of error in their measures. At this point, modeling was introduced as a process for explaining how the outcomes observed in the sample distribution were generated, as well as how other outcomes (measurements) like them might be generated. To initiate modeling, the teacher asked students to use TinkerPlots to create a model of "perfect measurement." This instruction was intended to draw students' attention to the fit between a model's simulated outcomes and the empirical sample students had previously constructed. When students observed the lack of fit between model and data, they began to list potential sources of error, such as inadvertent, random slips made when iterating a ruler. Students determined magnitudes and range in values for each proposed source of error. These values were typically established by conducting informal investigations (e.g., repeatedly leaving a small space or gap when iterating the ruler and determining the maximum magnitudes of under-estimation that typically resulted from this source). Figure 1 displays a facsimile of a student model in which chance slips in iterating a ruler might lead to under-estimates (i.e., gaps between iterations of the ruler, so that some length was left unmeasured) or overestimates (i.e., overlapping units on the ruler, so that some space was measured more than once). This model reflects a belief that errors of large magnitude are less likely than smaller magnitude errors. Students simulated the outcomes of the measurement process as a composition (sum) of signal, usually estimated by the classroom sample median measure of the perimeter, and the outcomes of the random devices, which represented different sources of error, such as those depicted in Figure 1.
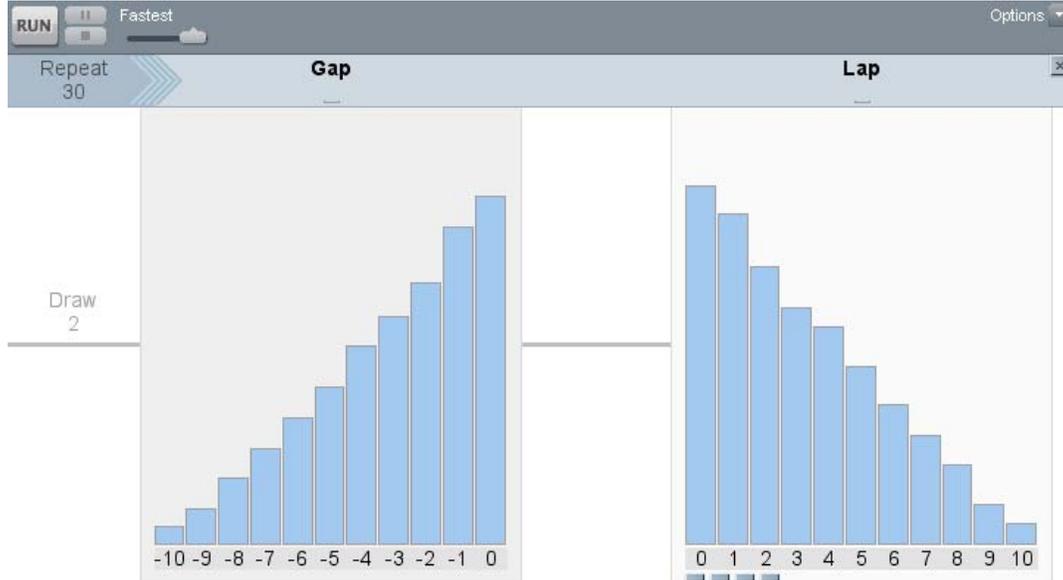
Figure 1. Student model of relative chance of different sources and magnitudes of iteration error

Establishing criteria for "good" models As students developed, compared, and contrasted their models, they discussed how to decide whether or not their model was good. A common initial approach was to compare perceptually the extent to which a simulated sample approximated the center clump of the sample. Students also tended to focus on the extent to which simulated sample statistics matched or came "near" to those of the sample. Some students noticed that as they ran the model more often, their initial impression of goodness was often misleading, perhaps because a sample-simulated sample correspondence was just "lucky." The teacher seized upon this notion of luck to remind students that the empirical sample is but one of many that they could have obtained, just as one simulated sample was one of many that the model could generate. Subsequently, most students collected sample statistics obtained from repeated runs of their models and constructed sampling distributions of their model's statistics. To assess model fit, they compared the empirical sample statistics of center and precision of measure to these sampling distributions. Decisions about fit, in turn, guided revisions to their models.

Using model-based sampling distributions to guide inference After deciding that their model was a good representation of the measurement process, students were shown new sample statistics purportedly constructed by another class, first a median and then an IQR. Students considered the possibility, proposed by the teacher, that the new measurements had perhaps been made on a table with a different perimeter. They also evaluated the claim that a change in the method of measure (e.g., by using a different measuring tool) made the measurements more precise. Students used TinkerPlots to partition their model-based sampling distributions and identified the percent of simulated medians or simulated IQR's that were the same as (or that exceeded or were less than) the new sample statistics. This process of model invention, revision, and model-guided inference was extended to other contexts of signal-noise, and

eventually, to natural variation. In all, 14 lessons were devoted to student investigations of sampling, constructing and revising models, and making inferences.

2. METHODS

2.1. PARTICIPANTS

Twenty-eight students in a Grade 6 class participated in all phases of the instruction. Of these, a convenience sample of 12 (modal age 11 years), consisting of 7 males and 5 females, was individually interviewed at the end of instruction. Interviewees were chosen on the basis of availability, given end-of-year schedules, and all had participated in most of the instructional activities within the design study. The middle school was located in the southern region of the United States, and at the time of the study, approximately 72% of the students qualified for lunch supplements (indicating comparatively low parental/custodial income).

2.2. PROCEDURE

Each student was interviewed either by the author or by a graduate assistant. Interviews were conducted in a separate room and were video recorded and later transcribed for analysis. The interview was semi-structured and flexible; all students responded to the same scenarios (described next) but interviewers were given flexibility in responding to each student's interpretation of a scenario. Consequently, although all students encountered a core set of questions, the interviewer often employed follow-up probes to obtain insight about how particular students were reasoning. Thus, the interview was a form of conversation intended to follow the contours of a student's reasoning (Ginsburg, Jacobs, & Lopez, 1998).

2.3. INTERVIEW PROTOCOL

The flexible interview consisted of three scenarios with related questions.

Scenario 1: Mystery spinner, sampling distribution The first scenario presented a "mystery" spinner in TinkerPlots. That is, the structure of the spinner was hidden (it appeared as a blank), but students were told that it was partitioned into 2-colors (in fact, the spinner area was 60% red, 40% blue). We asked students to speculate about the structure of the spinner and, if they said they were uncertain, to suggest the number of times that the spinner should be run to make a judgment, and to explain why. The purpose was to probe students' understanding of the value of a larger sample for inferring the structure of the device. These questions were directed toward students' conceptions of sample variability and sample size (Research Question 1).

We did not allow a student to run the spinner a large number of times, but instead ran it once 10 times. We then asked the student if s/he would like to guess about the structure of the spinner in light of its outcomes, or if, instead, s/he preferred to collect more samples of ten spins each (a sample size of 10). To probe conceptions of sample-to-sample variation, we asked students to anticipate what would happen in the next sample. We also asked them to justify the number of samples they considered necessary to be "more certain" about their predictions of the spinner structure.

Figure 2 displays an empirical approximation to a sampling distribution (500 samples) of a statistic (percent red), with the median and IQR of the sampling distribution visible. The interviewer showed the student this display and asked him or her to interpret a few of the cases on it (i.e., the interviewer pointed to a case value of 100%, a case value of 40%). Each student then inferred the structure of the spinner and interpreted the meanings of the median and IQR of the sampling distribution. Students predicted the effects on the median and IQR of a change in the sample size. We asked students to explain why a fictive observer noticed that, with a sample size of 10, a sample statistic of 10 percent red was rare in the sampling distribution, but this outcome was never observed with a sample size of 100. (Students did not actually view the second sampling distribution.) This sequence of questions informed the second research question.
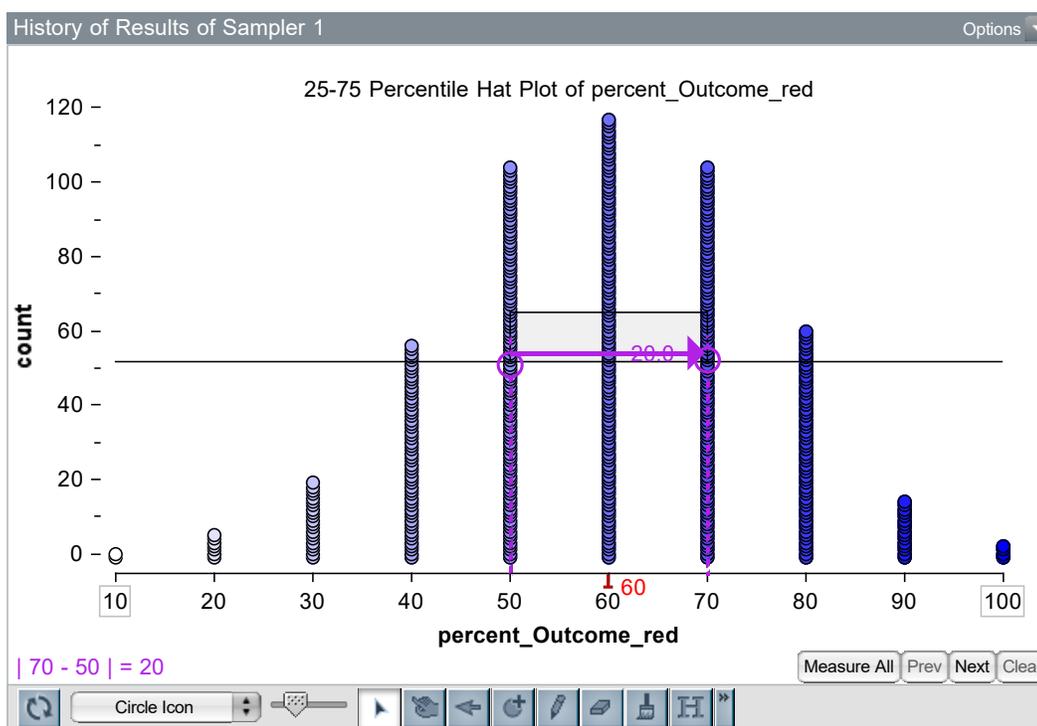

Figure 2. Sampling distribution of statistic, percent red, for mystery spinner

Scenario 2: Modeling skills, grounds of informal inference The second scenario involved a model of signal and noise. It was intended to elicit students' conceptions of models and to assess their practical modeling skills. The scenario also provided an opportunity to ask about students' criteria for deciding whether the model presented was "good." Following this exploration of student conceptions and skills, students saw sampling distributions of sample statistics generated by the model and then made and justified a decision about whether or not a statistic obtained from a new empirical sample represented a real change. This sequence of questions informed the third and fourth research questions.

Students were informed first of the results of an analysis of the magnitudes and likelihood of errors that resulted in an over- and under-estimation of the circumference of a person's head.

The errors were described as resulting from using a tape measure. The interviewer explained that sometimes the measurer pulled the tape too tightly, so that it stretched, and on other occasions too loosely, leaving slack in the tape. The analysis was represented by a model that combined the sample median and tape error, as illustrated in Figure 3. The model was simpler than those invented by students during instruction.
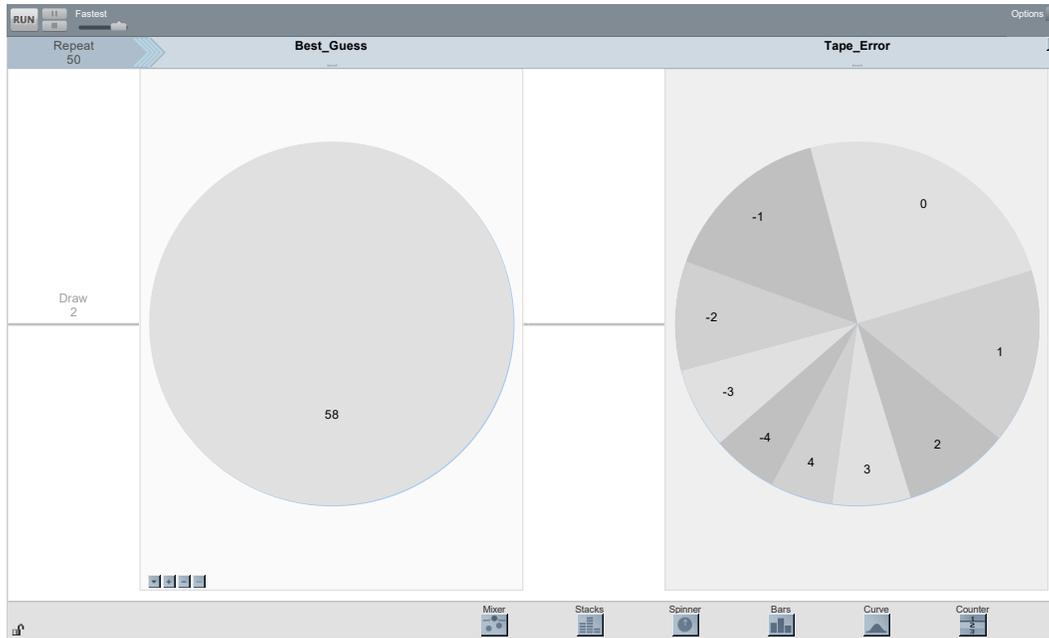

Figure 3. Model representing measurements of head circumference

Students predicted the shape of the data that would result from running the model. Students also described the criteria they considered when deciding whether or not a model was a "good one." After these initial responses, students were shown a sample of the "actual measurements," with statistics of median and IQR visible (Figure 4). They were also invited to run the model to decide whether this particular model was a good one. If the student did not bring it up, the interviewer asked whether one run or multiple runs of the model would be better for informing this decision, and to explain why they thought so.
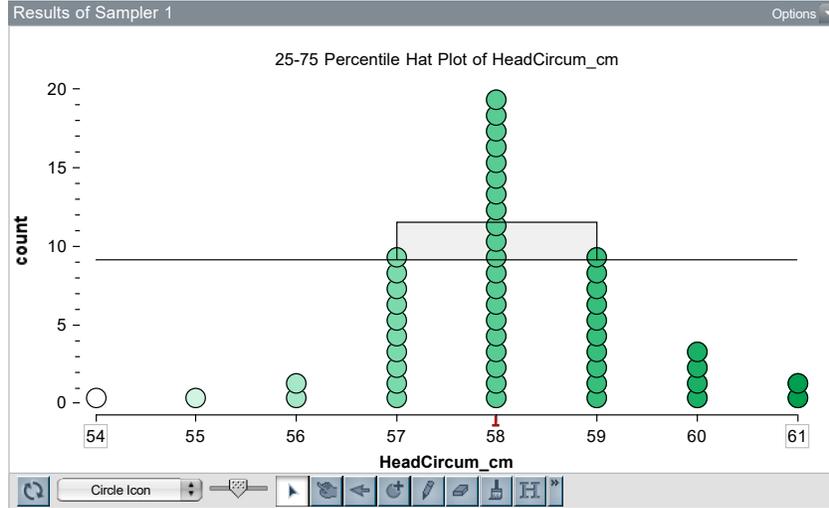
Figure 4. Sample distribution of head circumference measurement

After recording students' impressions and investigations with the model, we presented two empirical sampling distributions of model parameters (n = 200 simulated samples of size 50), one of the simulated median and the other of the simulated IQR (see Figure 5). The sampling distribution of medians was tightly centered about the sample median of actual measurements, but the median IQR of the sampling distribution of simulated IQR was 3, not 2, as obtained in the actual sample. To assess generalization of the meaning of a case in a sampling distribution, we asked students about the meaning of the case values in the sampling distributions (repeating the question posed in the first scenario, albeit now about simulated samples generated by the model). The interviewer asked again about student impressions of the goodness of the model in light of the sampling distributions.
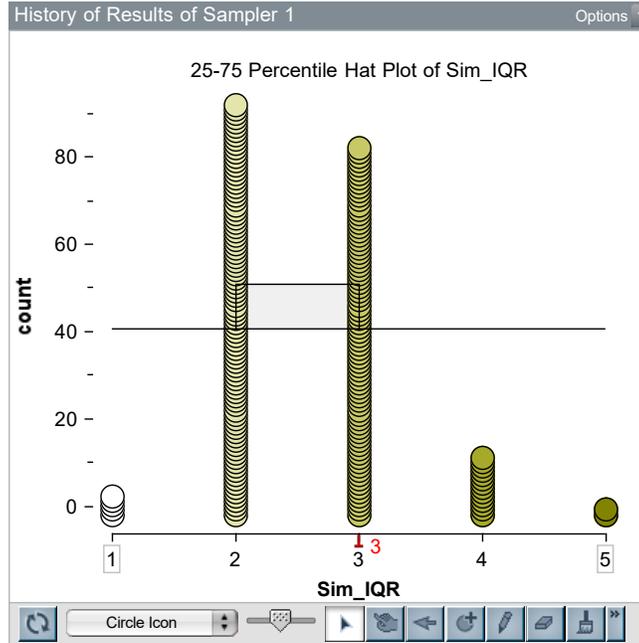
Figure 5. Sampling distribution of simulated sample IQR's

Students then considered a different sample median of 60 cm., described as obtained by another group of measurers. We asked students to decide if the new median represented the same person's head or the head of a different person. Students explained their decision. Last, we asked students how they would alter the model to produce a U-shaped sample distribution, and then, how to alter the model to produce a uniform distribution.

Scenario 3: Modeling production, informal inference The third scenario addressed students' conceptions of model-based inference in the context of a process for manufacturing batteries. It was designed to inform the fourth research question. Students first interpreted a dot plot of the lifetime (in minutes) of 50 batteries, and responded to interviewer questions about the likely target value of the process and the consistency of the products. (The median and IQR were visible in the display).

The interviewer then revealed a "good" model of the production process that included sources of variability that affected battery life, such as chance variations in chemical composition and temperature, as displayed in Figure 6. Students also saw model-based empirical sampling distributions of simulated sample median (see Figure 7) and IQR for batches of 50 batteries. For each sampling distribution, the interviewer asked students about the likelihood that a sample median or an IQR was equal to or exceeded a particular value just by chance. The intention was to assess students' skills in relating the sampling distribution to probability.
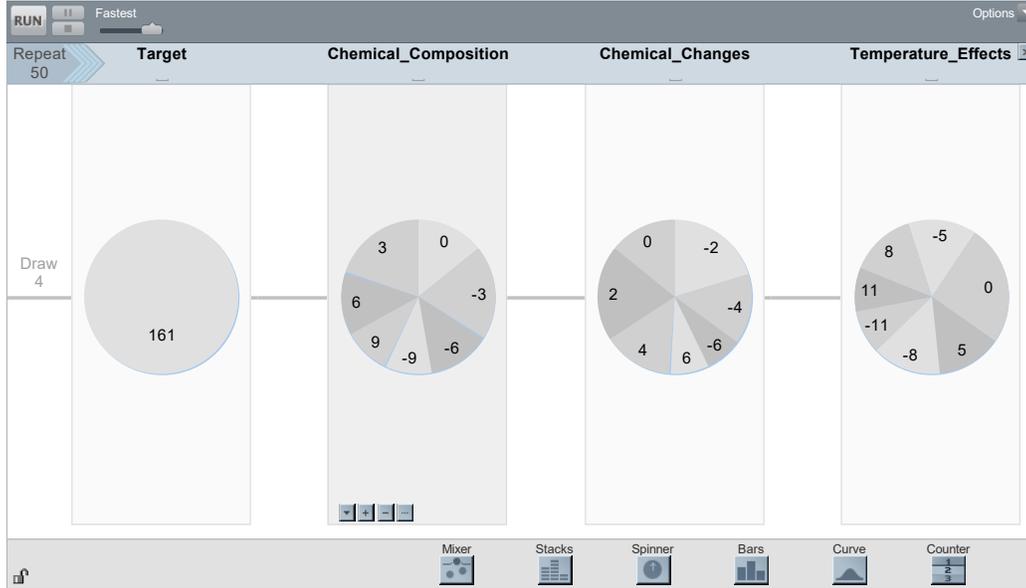
Figure 6. Model of production process for batteries with lifespan in minutes
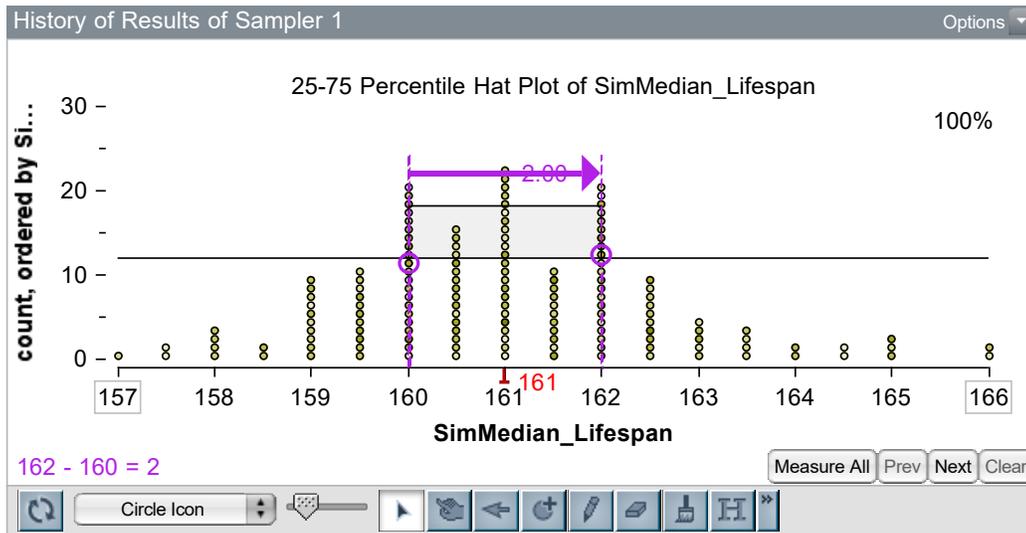

Figure 7. Sampling distribution of median for 100 simulated samples of size 50

Students were informed that the battery company had financed a new production process to increase the lifespan and consistency of their product. The company tried out this process to produce a batch of 50 new batteries, and the head of research claimed that the new production median of 164 represented a real improvement over the previous process. Students were asked how the head of research might use the sampling distribution obtained with the model of the old process to buttress this claim. About 9% of the sampling distribution of model-simulated medians met or exceeded this value. Students were asked how someone might respond if he

were less convinced that the new process was a genuine improvement. Students made a similar determination about a claim that the new process resulted in improved consistency of battery lifespan, based on a sample IQR of 8 for the batch of 50 new process batteries. In the sampling distribution of model-simulated IQR's, this value occurred rarely (1%).

Coding student responses To analyze the interviews, we adapted a coding scheme developed in a previous iteration of the design that was also conducted with sixth-grade students (Lehrer, Jones, & Kim, 2014). The initial scheme was carefully reviewed by a visiting scholar in statistics education, who carefully reanalyzed two of the video recorded interviews conducted during this investigation and identified a few distinctions that were not included in the initial scheme. The interview coding protocol was adapted to reflect these novel elements.

Annotated transcriptions were developed for each interview, including interviewer questions and probes, student responses, interpretations of elements of student knowledge, and additional notes about possible ambiguities in interpretation or additional comments about the potential significance of student talk or gesture. I then coded all transcripts, and the visiting scholar recoded half of the them. Disagreements were rare and were resolved by consensus. I inspected the codes across scenarios to create analytic themes that were responsive to the research aims of this inquiry, following a grounded inquiry approach (Corbin & Strauss, 2008).

3. RESULTS

3.1. CONCEPTIONS OF SAMPLE AND SAMPLING VARIABILITY

Recall that the first interview scenario involved guessing the hidden structure of a 2-color spinner. Student responses to questions about this scenario revealed how they conceived of sample variability and how they thought about the influences of sample size on this type of variability. Students' interpretations of a sampling distribution of a statistic generated for each sample (percent red) indicated how students conceived of cases in a sampling distribution and how they predicted and explained the influence of sample size on statistics of center and variability.

Sample variability All but one student (n =11, 92%) initially were uncertain about the hidden structure (e.g., "It could be anything"). The exceptional student, CA (initials of pseudonymous identity), said that he knew the spinner structure was 50-50 and further suggested that two repetitions of the spinner would verify his prediction. In contrast, all the others reported that observing the outcomes of a large number of repetitions of the process would increase their confidence in an estimate of the percentage of the spinner area that was red. For example, JP mentioned that "…the more points (outomes) you have, the closer it gets…it will round out to what the real spinner is." JP went on to make an analogy to a related random process of coin flipping: "If you flip a coin a billion times, the decimals will round out to… [gestures 50-50] and [be] a good estimate." Another student, GB, said, "If you only had a few (repetitions of the spinner) you would only be able to see what happened with those few, but if you have many, you could see what happens with the big picture." A third, SD, imagined a very thin slice of one color and said one would need a very large number of repetitions, "…because the smaller it is (the sector occupied by one color), the more times you have to repeat it, just in case." Some students (n =5, 42%) further explained their confidence in large

samples by claiming that increasing "precision," a statistic of the variability of measurements invented during instruction, is associated with increasing sample size. For example, TK declared, "The more you run it, the more precise it's going to be," and CS said, "Find out its structure by running it and running it again…and again, enough so that I can see more of a precision" [This statement was accompanied by a hand gesture illustrating a decreasing span]. Yet another student, IW, felt that running the spinner a large number of times would result in a more precise estimate. She contrasted her preference for a large sample to a sample of size 10: "…because it's just by chance, so like, if it's a 50-50 spinner, then it could be like 20% red, 80% blue" [with a small sample]. She went on to point out that with a larger sample, the estimate would more likely approximate the true ratio. Students who did not explicitly invoke precision nevertheless tended to associate more outcomes with increasing the likelihood of more "accurate" estimates of the structure of the spinner. This language of "accuracy" appeared to reflect an expected reduction in the variability of estimates with increased sample size (here explicitly meaning number of repetitions of the process). In summary, for this simple process, students appeared to hold an image of a sample as originating in the generation of variable outcomes resulting from repetition of a chance process. Moreover, nearly all the students appreciated that a statistic estimated a population parameter (e.g., percent red area of a spinner) better with large samples rather than with small ones. Students anticipated a reduction in variability of estimated values with increasing sample size.

Sampling variability Given a constraint of a sample size of 10, all students preferred a larger, rather than smaller, number of samples to guide inference. The previous exceptional student, CA, justified his switch in thinking as follows: "Because with a large number of samples, you can get it (estimate of percent red) more precise." Every student interpreted each case of the sampling distribution as a statistic that summarized a sample of 10 repetitions of the mystery spinner, with the statistic indicating the percentage of the 10 repetitions that were the indicated outcome (e.g., those landing in the red sector). For example, CA noted that a case value of 40% red in the sampling distribution meant: "Based on this [indicating a case value in the sampling distribution], 4 [outcomes] were red, 6 were blue one time out of the 500 (samples) collected."

All students judged that the median estimated the percentage of red in the mystery spinner (e.g., LL: "Like, with the median is at 60, you are not going to put, you are not going to have, be guessing that the spinner would be 50-50"). For students, the sensibility of the median as a valid estimate was buttressed by its location within the center clump of the sampling distribution. For example, BP said, "The data is mainly right here (gestures to center clump); it looks like the center to me, and 60 percent is the median." Considering the influence of a change in sample size from 10 to 100, all students predicted that the median of the sampling distribution would not change. The students explained their anticipation of no or little change either by appealing to the invariant structure of the spinner (e.g., GB, "because you are still keeping it 60-40"; SD, "because you're not changing the spinner you're just changing how many times it repeats") or by noting the increased precision of estimate for larger samples. For example, CA explained that the median would "…stay the same because you made the sample size 100. That will mean that the data will get more precise and more clumped up and that will mean that the median here [gestures to column of values that include the median] will stack up more on the 60."

All but one student (92%) predicted that the increase in sample size from 10 to 100 would decrease the IQR of the sampling distribution. Students' justifications typically included an expectation of increased precision with larger sample sizes and a corresponding concentration of cases about the center. For example, CA explained, "I think it [the IQR] will decrease, because if you take away (gestures to both tails of n=10 sampling distribution while simulaneously decreasing his hand span) since we run it more times, it will be more precise, so then these would decrease (gestures toward tails) and these (center values) will get more clumped, which means that the 50 percent range will get smaller."

At the conclusion of this portion of the interview, students explained the absence of extreme cases (i.e., 10% percent-red) in a sampling distribution with a sample size of 100 and the presence of such cases in a sampling distribution with a sample of size 10. Students' justifications (92%) were consistent with their previous interpretations that invoked the unchanging structure of the spinner. For example, GB said, "It [sampling distribution of sample size 100] would have to be exactly 10 out of those 100 times; that's a lot harder to do than just 1 out of 10, because the odds are against (that) red percentage. It's not made to be—it's not supposed to be less than what it is, which is 60%."

In summary, in this simple scenario involving a hidden spinner, students appeared to construct a hierarchical image of sample with sample statistics understood as case values of a sampling distribution and with statistics of the sampling distribution interpreted as estimating characteristics of the generating process. In the next sections, students' conceptions of sample are revisited in the more complex context of modeling processes that include multiple components of signal and noise.

3.2. STUDENTS' MODELING SKILLS AND CRITERIA FOR "GOOD" MODEL

To employ models as tools for thought, students must be able to conceive of relations between components of the model and distributions of outcomes, and they must develop criteria that inform their judgment of the adequacy of a model. The second scenario, which involved repeated measures, provided a window to these aspects of student thinking.

Modeling skills All students (n =12) modified the model of signal and noise appropriately to produce alterations in the distribution of outcomes by revising the proportions of error that would result in either more uniform or U-like distributions. For example, to create a uniform distribution, CS pointed to the error component of the model of head circumference measurment depicted in Figure 3 and indicated how a more uniform distribution could be obtained: "Just equalize angles (a TinkerPlots function). That way they would be all equal (each magnitude of error) so we would have an equal chance of getting 54 thru 62 (the sum of the fixed signal and random error). And so to do that we need to alter all of them to make them all equal and add up to 100 and make it as likely as we can" (Gestures flat distribution with hand moving level above table). For the bi-modal, U distribution, CS again looked at the area occupied by each magnitude (e.g, +4 or -4) of the model of head circumference. He suggested that: "We'd have to reverse everything pretty much…so I think -1, 0 and 1 would get smaller while -2, -3, -4, 4, 3, and 2 will get bigger (with the largest magnitudes of error now occupying the largest proportion of the area of the spinner.) Students had not practiced this particular problem during instruction. Hence, their performance indicates that they were able to imagine

how particular configurations of magnitude and likelihood of random errors might influence the distribution of outcomes.

Criteria for good models When the interviewer asked students how they decided whether a model was a good one (e.g., "When you think about whether or not a model is good, what kinds of things do you consider?"), all students suggested multiple criteria. A majority (75%) advocated that a model should explain or represent a process. For example, LL suggested that a good model is one that "…shows what actually happened and if they count in all the errors. And how much those errors weighed in on it." All students stated that good models generated outcomes that approximated those of empirical samples. As QC stated, "Because there's always chance and it can always be different. But it can't be too different." The idea of models as approximations was often expressed as reference to "possible values" generated by the model but not present in the empirical sample. JH recalled a time during instruction "…when there was no 9 on the graph [the empirical sample value] but it was possible, so you still have to account for it [in your model]." The grounds of approximation typically included statistics of center and variability and the shape of the data. As AC comprehensively summarized, "See if the measures of center are close together, take the shape of the data, also see your endpoints, range, IQR, and average deviation."

However, students suggested that comparisons such as those suggested by AC should be based on generating many simulated samples with the model, not a single simulated sample. As LL noted, "…then you know if your model is being consistent." Half of the students (n =6) spontaneously mentioned the value of generating sampling distributions of model statistics to compare with sample statistics, a feature supported by TinkerPlots. For example, CS said, "Also a tactic that we use that you might want to do is to collect statistics." When presented with sampling distributions of model statistics, all students followed through on this strategy, making judgments about the goodness of the model displayed in Figure 2 by comparing the sampling distributions of model statistics with their counterparts in the empirical sample. In summary, student responses to the second scenario indicated that they had developed modeling competencies, a common interpretation of models as approximations of processes, and a wide range of strategies and criteria for judging model fit. Of these, use of sampling distributions indicated a model-based generator of a hierarchical image of sample. In students' views, possible values produced by models generated sample-to-sample variability, and this variability was encapsulated by sampling distributions of model-based statistics.

3.3. STUDENTS' MODEL-BASED INFORMAL INFERENCE

The influences of models and model-based sampling distributions on students' thinking about the grounds of inference are next described for repeated measure (scenario 2) and production (scenario 3) processes. In each scenario, assuming a model of a process was valid, students decided about the validity of claims made about alterations to the modeled process on the basis of a statistic that represented a sample purportedly drawn from the altered process.

In the repeated measure (second) scenario, students responded to a question about whether a new sample median of 60 might indicate measures of a different person. Half of the students spontaneously employed the model-based sampling distribution to make this inference. For example, GB decided, "Based on this [sampling distribution of simulated model medians], 60 [the new sample median] isn't even a possibility. But at the same time, there could be a chance

that it could be 60 because they could have messed up in some way." The interviewer asked GB to clarify, and he answered that even though the value of 60 was absent in the sampling distribution of 200 samples, it might still appear with many more samples, making it "…possible, just not likely." When asked for further explanation of their decisions, two other students also referred to the sampling distribution. The other four students justified their inferences by superimposing an imagined translation of the existing sample distribution centered about the new sample median. They judged that a center clump shift of this magnitude was not likely under conditions of repeated sampling and concluded that the new sample consisted of measures of a different individual. For example, AC decided in this way: "[It's] a different person, becasuse if they did measure the same person, I think their numbers would match up pretty closely with this one [the existing sample]. Not exactly, but pretty closely, in this, it [the new sample median value in the original sample] only occurred 4 times out of 50, that's almost, like 1/10 and their median [the new sample] was 60, so I think they measured a different person."

In the third signal/noise scenario about the production of batteries, the interviewer and student first established that the median statistic of the sample estimated the likely target value of the production, and that the sample IQR was a measure of the consistency of the product. Recall that students had access to an empirical sample of battery life spans, a model of the manufacturing process with components consisting of a visible target value and different sources of error, and sampling distributions of the simulated model medians and IQRs. With this information, students were asked how they would support the claim, made by a fictive head of research, that the new production process resulted in a real improvement in average duration of batteries. Students also considered how a skeptic, who did not believe the head of research, could use the same data to conclude that no real improvement had been achieved.

Students inspected a sample median representing the improved process. Ten of the 12 students (83%) referred to the sampling distribution of model-simulated median battery life span to make a decision. For example, BP justified his decision that average life span had improved by pointing to the model-generated simulated sampling distribution of medians, declaring, "It does not look like it is very likely, there are only 9 dots in there [counting the values in the sampling distribution of 164 and higher], 9% likely." He went on to suggest that the skeptic might "…look at that (the same region) and say that it's a pretty big portion." The other two students imagined a shift in the center clump of the sample consistent with the new value of 164, but deemed it unlikely just by chance. They thus considered the shift as supporting the claim of real improvement. With regard to the skeptic (the counterfactual), 4 of the 11 who were asked mentioned a pragmatic concern—the small magnitude of the difference probably would not be noticed by consumers. Seven other students reflected BP's perspective and cited the sampling distribution as indicating that the median value did occur in that distribution, so "…there could be a little chance, but it's probably not."

When asked whether product consistency had improved, all but one student employed the model-simulated sampling distribution of the IQRs to decide that the new sample's IQR was not likely to have been generated by the old production process. For example, BP mentioned that an IQR of 8 appeared only once in the sampling distribution. "That's 1%, and with the [new empirical sample representing the improved process] you got an IQR of 8 immediately and that's not very likely." Similarly, JH imagined the sampling distributions of the modeled (old) and improved (new) processes, saying, "Well, if our only sample is 8 [the new process empirical sample IQR], this one [the model-simulated sampling distribution of IQR] says it

only happens one time out of 100 times. So, this one [again pointing to model-generated sampling distribution], I'm guessing that their first sample was a 14 [imagines the process of constructing the simulated sampling distribution] because that is what happened the most, so if our first sample [the improved process] is an 8, then I think it's going to happen more times." In light of the duration of the interview, we asked only half of the students for counter-arguments to this second claim of a more consistent product. Five of these students again referred to the sampling distribution to justify skepticism (e.g., GB: "This is just one time (sample)..it's not likely, (but) it's possible").

In summary, the use of modeling to generate images of sample-to-sample variability was leveraged by these students to extend the reach of informal inference. When students were asked to judge the validity of a model of a process that involved tangible components of random noise and signal, many students' inferences were guided by a hierarchical image of sample supported by sampling distributions of model statistics. JH's responses perhaps best exemplify a network of coordinations among modeling, hierarchic image of sample as represented by the sampling distribution of model-generated statistics, and inference.

4. DISCUSSION

During eight weeks of instruction, students had repeated opportunities to experience and conceptualize multiple aspects of variability, to model variability by generating processes involving signal and noise, and ultimately, to consider how inference could be sustained in light of chance variation. The findings of the interviews and previous research support a possible pathway for introducing young students to model-based inference by leveraging contexts of signal and noise (Konold & Lehrer, 2008).

The early phases of instruction emphasized describing, representing, and measuring variability, practices that served as a foundation for getting a grip on variability (Reading & Reid, 2010; Reading & Shaughnessy, 2004; Wild & Pfannkuch, 1999). Moreover, because the initial measurement task was tangible and because multiple sources of variability were evident to students via their firsthand generation of measurements, students could construct an image of variability as arising from a repeated process. Understanding how variation arises (Pfannkuch & Wild, 2004) and building an image of repeated process are seeds of statistical inference (Thompson et al., 2007). Students' interpretations of measurement as composed of true measure (signal) and "mistakes" (errors) helped them reconcile expectation and variability, concepts that are further bedrock for understanding variability (Konold & Pollatsek, 2002; Petrosino, Lehrer, & Schauble, 2003; Watson, Callingham, & Kelly, 2007). These initial foundations were likely important for students' subsequent explorations of modeling these processes.

Unlike operating random devices (e.g., spinners, dice), which students readily attribute to chance, modeling requires the additional representational step of imagining how the syntax of a process that does not ostensibly involve the operation of chance devices can nevertheless be well represented by them. This is typically a difficult step for young people (Lehrer & Schauble, 2006), that requires going beyond just achieving a firm grasp of the behavior of chance devices. In this instructional sequence, students drew upon statistics they originally developed to measure signal and noise (e.g., precision) of repeated measures to understand better the outcomes generated by random devices. As suggested by their responses to the first scenario, which involved a hidden random device, students tended to interpret the operation of

the random device as governed by a signal, the structure of the spinner, and chance. They expected increasing sample size to increase the precison of estimates of the structure of the spinner. Student interpretations of sample-to-sample variation in the operation of the random device were similarly guided by conceptions of precision and signal (structure). Hence, they predicted that increases in sample size would not influence a statistic of center of the sampling distribution, but would increase the precision of estimation (and, accordingly, decrease a statistic of variability).

Employing these simple random devices, students invented and revised models of measurement and production processes. These models were composed of signal and random errors that were represented by multiple random devices. As they ran their models, students came to understand models as approximations to processes, rather than as direct copies. This perspective of model approximation was a gateway to reasoning about any sample as an instance of a succession of samples—each trial of the model generated what students conceived of as possible values, and the collection of these values constituted a simulated sample. Conceiving of model outputs as possible values helped students come to see any particular sample, even one they had generated, as only one of many potential samples. These discoveries thus supported their development of a hierarchical image of sample (Saldanha & Thompson, 2002, 2014).

Sampling distributions of model statistics further expanded students' grasp of inference in light of uncertainty. One expansion was that sampling distributions were accepted as a conceptual tool for judging model fit, as students began to recognize that one or two instances of correspondence between a simulated sample and a sample in the world could occur just by chance. Consequently, students' criteria for "closeness" when guaging model fit shifted to include attending to the center clump of sampling distributions of simulated statistics and to their correspondence with the center and variability of the empirical sample. A second extension was students' use of sampling distributions to evaluate claims about changes in a process. With sampling distributions, students had a tool for quantifying the variability of a statistic from sample to sample. They did so by locating the value of a new sample statistic within the sampling distribution of the statistic generated by a "good" model.

Most students appreciated that decisions made in light of the bands of uncertainty suggested by the sampling distribution of model statistics could be mistaken. That is, they understood that what was at stake was chance, rather than certainty. Manor Braham and Ben-Zvi (2015) and Manor Braham (2016) described this form of thinking as probabilistic reasoning about sampling distributions. Even those students who did not explicitly employ the sampling distribution to guide their inference nonetheless judged the plausbility of a claim by imagining the likelihood of claims made about sample statistics in light of sampling variability about the existing sample (Lehrer & Schauble, 2004). That is, if one imagined a claimed statistic as a new center of distribution, how likely would such a shift be? This is a form of distributional thinking that was likely supported by conducting multiple runs of a model and observing the resulting variability in outcomes.

The instructional approach taken here shares a commitment with other programs of research that employ modeling to generate sampling distributions as a way of expanding young students' grasp of inference. The properties of contexts that afford productive conceptual challenges and opportunities as students invent and revise models deserve further study. For example, Konold and colleagues situated inference in modeling processes that involve signal and noise (e.g., Konold & Harradine, 2014), whereas Manor Brahm and Ben-Zvi (2015)

supported students in generating sampling distributions by re-sampling questionnaire data about personal preferences that students and their peers have generated. In the research conducted by Manor Brahm and Ben-Zvi, students determined the size of a sample required to achieve sufficient confidence about an inference relating to the questionnaire data. These two approaches may have different pedagogical advantages. In the context of signal and noise, processes can be experienced firsthand, whereas those involved in individual differences and natural variability are largely invisible (Lehrer & Schauble, 2007; Konold & Lehrer, 2008). For example, personal preferences likely arise from complicated histories. Yet personal perferences may be of more sustained interest to students than signal-noise processes of measuring or making products. Perhaps learning environments should deliberately span multiple forms and contexts of modeling in which probabilistic inference is at stake.

The limitations of this research imply that its conclusions are tentative. The process of model-based inference unfolded as the culmination and coordination of related practices of visualizing and measuring variability, which are important building blocks of modeling. The classroom was led by a skillful teacher who was attuned to students' thinking. The sample was restricted to convenience due to students' schedules. And there is no evidence concerning whether these young students would routinely anticipate the generation of a sampling distribution for any process for which they had reasonable expectations that chance was operating. It is also worth reiterating that most of the modeling conducted by students was of systems and processes that were visible, contexts in which they could causally influence the distribution of outcomes (e.g., by using different tools to measure, or by changing methods of making products). This task feature has previously been acknowledged as an important support for grasping variability (e.g., Biehler, 1999; Pfannkuch, 2011). Perhaps most critically, modeling variability initatiated these students into the dialog between data and models that Cobb and Moore (1997) describe as essential to statistical practice (Pfannkuch & Wild, 2000). It is this integration between data generation and modeling that will likely prove most fruitful for students in the long run.

ACKNOWLEDGEMENTS

REFERENCES

Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal*, *3*, 64–83. Retrieved from http://iase-web.org/documents/SERJ/SERJ3(2)_Bakker.pdf

Ben-Zvi, D., Bakker, A., & Makar, K. (2015). Learning to reason from samples. *Educational Studies in Mathematics, 88*, 291-303.

Biehler, R. (1999). Learning to think statistically and to cope with variation. *International Statistical Review*, *67*(3), 259-262.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, *103*(9), 801-823.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, *32*(1), 9-13.

Corbin, J., & Strauss, A. (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Thousand Oaks, CA: Sage.

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning*. New York: Springer.

Ginsburg, H. P., Jacobs, S. F., & Lopez, L. S. (1998). *The teacher's guide to flexible interviewing in the   classroom. Learning what children know about math*. Boston: Allyn and Bacon.

Konold, C. (2007). Designing a data analysis tool for learners. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp.267-291). New York: Lawrence Erlbaum Associates, Taylor and Francis.

Konold, C., & Harradine, A. (2014). Contexts for highlighting signal and noise. In T. Wassong, D. Frischemeier, P. R. Fischer, R. Hochmuth, & P. Bender (Eds.), *Mit Werkzeugen Mathematik und Stochastik lernen: Using Tools for Learning Mathematics and Statistics* (pp. 237-250). Wiesbaden, Germany: Springer.

Konold, C., & Lehrer, R. (2008). Technology and mathematics education: An essay in honor of Jim Kaput. In L. D. English (Ed.), *Handbook of International Research in Mathematics Education* (2nd ed., pp. 49-72). Philadelphia: Taylor & Francis.

Konold, C. & Miller, C. D. (2011). *TinkerPlots: Dynamic data exploration.* Emeryville, CA: Key Curriculum Press. Available from www.tinkerplots.com

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, *33*(4), 259-289.

Lehrer, R., Jones, S., & Kim, M-J. (2014, April). *Model-based informal inference*. Paper presented at the Annual Research Meeting of National Council of Teachers of Mathematics, New Orleans, Louisiana.

Lehrer, R., & Kim, M. J. (2009). Structuring variability by negotiating its measure. *Mathematics Education Research Journal*, *21*, 116-133.

Lehrer, R., Kim, M-J., Ayers, E., & Wilson, M. (2014). Toward establishing a learning progression to support the development of statistical reasoning. In J. Confrey & A. Maloney (Eds.), *Learning over time: Learning trajectories in mathematics education* (pp. 31-59). Charlotte, NC: Information Age Publishers.

Lehrer, R., Kim, M-J., & Jones, R. (2011). Developing conceptions of statistics by designing measures of distribution. *International Journal of Mathematics Education*, *43*(5), 723-736.

Lehrer, R., Kim, M., & Schauble, L. (2007). Supporting the development of conceptions of statistics by engaging students in modeling and measuring variability. *International Journal of Computers for Mathematics Learning*, *12*, 195-216.

Lehrer, R., & Schauble, L. (Eds). (2002). *Investigating real data in the classroom*. New York: Teachers College Press.

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

Lehrer, R., & Schauble, L. (2006). Cultivating model-based reasoning in science education. In R. KeithSawyer (Ed.), *Cambridge Handbook of the Learning Sciences* (pp. 371-387). Cambridge: Cambridge University Press.

Lehrer, R., & Schauble, L. (2007). Contrasting emerging conceptions of distribution in contexts of error and natural variation. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp.149-176). New York: Lawrence Erlbaum Associates, Taylor and Francis.

Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning*, *13*, 152-173.

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, *8*(1), 82-105. Retrieved from http://iase-web.org/documents/SERJ/SERJ8(1)_Makar_Rubin.pdf

Manor Braham, H. (2016). *Developing elementary school students' reasoning with samples and sampling using the integrated modeling approach*. Doctoral Dissertation, University of Haifa, Israel.

Manor Braham, H., & Ben-Zvi, D. (2015). Students' articulations of uncertainty in informally exploring sampling distributions. In A. Zieffler & E. Fry (Eds.), *Reasoning about uncertainty: Learning and teaching informal inferential reasoning* (pp. 57-94). Minneapolis, MN: Catalyst Press.

McClain, K., & Cobb, P. (2001). Supporting students' ability to reason about data. *Educational Studies in Mathematics*, *45*, 103-129.

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.

Pfannkuch, M. (2011). The role of context in developing informal statistical inferential reasoning: A classroom study. *Mathematical Thinking and Learning*, *13*(1&2), 27-46.

Pfannkuch, M., & Wild, C. J. (2000). Statistical thinking and statistical practice: Themes gleaned from professional statisticians. *Statistical Science*, *15*(2), 132-152.

Pfannkuch, M., & Wild, C. (2004). Towards an understanding of statistical thinking. In D. Ben-Zvi, & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp.17–46). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Pratt, D., & Ainley, J. (2008). Introducing the special issue of informal inferential reasoning. *Statistics Education Research Journal*, *7*(2), 3-4. Retrieved from http://iase-web.org/documents/SERJ/SERJ7(2)_Pratt_Ainley.pdf

Prodromou, T., & Pratt, D. (2006). The role of causality in the coordination of two perspectives on distribution within a virtual simulation. *Statistics Education Research Journal*, *5*(2), 69-88. Retrieved from https://iase-web.org/documents/SERJ/SERJ5(2)_Prod_Pratt.pdf

Reading, C., & Reid, J. (2010). *Reasoning about variation: Rethinking theoretical frameworks to inform practice*. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society*. (Proceedings of the 8th International Conference on the Teaching of Statistics, Ljubljana, Slovenia, July). Voorburg, The Netherlands: International Statistical Institute. Retrieved from https://iase-web.org/documents/papers/icots8/ICOTS8_8E2_READING.pdf

Reading, C., & Shaughnessy, J. M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 201-226). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Rubin, A., Bruce, B., & Tenney, Y. (1990). Learning about sampling: Trouble at the core of statistics. In D. Vere-Jones (Ed.), *School and general issues* (Proceedings of the 3rd

International Conference on the Teaching of Statistics, Dunedin, New Zealand, August 19-24). Voorburg, The Netherlands: International Statistical Institute. Retrieved from https://iase-web.org/documents/papers/icots3/BOOK1/A9-4.pdf

Saldanha, L., & Thompson, P. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics*, *51*, 257-270.

Saldanha, L., & Thompson, P. W. (2014). Conceptual issues in understanding the inner logic of statistical inference: Insights from two teaching experiments. *Journal of Mathematical Behavior*, *35*, 1-30.

Thompson, P. W., Liu, Y., & Saldanha, L. (2007). Intricacies of statistical inference and teachers' understanding of them. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 207-231). New York: Lawrence Erlbaum Associates, Taylor & Francis.

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

Wild, C. J. (2006). The concept of distribution. *Statistics Education Research Journal*, *5*(2), 10-26. Retrieved from http://iase-web.org/documents/SERJ/SERJ5(2)_Wild.pdf

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

RICHARD LEHRER
Department of Teaching and Learning
Vanderbilt University's Peabody College
1930 South Drive
Nashville, TN