1. INTRODUCTION

Beginning in the mid 1980s, statistics-related activities began to appear with regularity in the United States pre-university curricula (Watkins, Burrill, Landwehr, & Scheaffer 1992). Influenced in part by the Quantitative Literacy Series (Landwehr 1985), the approach taken in those curricula increasingly drew on the works of John Tukey and his revolutionary approach, Exploratory Data Analysis (Tukey 1977; Biehler 1989).

The hope was that the informal tools and philosophy of Exploratory Data Analysis (EDA) would permit students to dive immediately in to the analysis of real data, to look for interesting patterns and trends without worrying about formulating and testing hypotheses. In this way, students could learn basic ideas of statistics without getting bogged down in formal statistical inference, with its reliance on probabilistic and counterfactual reasoning ("If there were really no difference in the two groups, the probability of this result would be x."). Accordingly, Watkins et al. (1992) recommended that introductory statistics at the secondary level begin with "basic ideas of data analysis, informal probability, simulation, and an intuitive introduction to statistical inference" (p. 49). In a similar spirit, David Moore (1992) argued that introductory statistics courses at the university level "should cover no more probability than is actually needed to grasp the statistical concepts that will be presented" (p. 23). To the implied question, "How much probability do younger students need to be able to grasp basic data-analysis concepts?" the de facto answer arrived at for elementary and middle school students in the United States was "none." Following the lead of the National Council of Teachers of Mathematics Standards documents (1989; 2000), existing middle school and elementary curricula treat the two topics as separate and independent strands.

As researchers began studying settings in which students were introduced to EDA, an unsettling picture soon emerged. Students given considerable exposure to and instruction in data-analysis techniques nevertheless had difficulties performing one of the most basic tasks in analyzing data — judging whether two groups appeared different by comparing their averages or the approximate centers of their distributions (Hancock, Kaput & Goldsmith 1992; Watson & Moritz 1999; Konold, Pollatsek, Well, & Gagnon 1997; Biehler 1997). This shortcoming has prompted researchers to delve more deeply into the underlying understandings and skills that would allow

students to make and justify such comparisons. Recently, these required understandings have been reconceived as a set of "informal inference" skills. Broadly speaking, informal inference involves making generalizations from samples without the use of formal statistical tests and the quantification of uncertainty that come with them (e.g., Makar & Rubin 2007). Looking across many of the descriptions of informal inference (e.g., Rubin, Hammerman, & Konold 2006; Pfannkuch 2006) we find repeated references to ideas associated with chance. It thus appears that many statistics educators are now taking seriously a possibility that Rolf Biehler suggested shortly after EDA found its way into classrooms: that many of the difficulties novices experience stem from "a too simple (probability-free) educational conception of data analysis" (1994, p. 20).

For the past three years, and with funding from the National Science Foundation, we have been developing tools and materials for teaching data analysis and chance in the middle school. The main objective of the project is to help middle school students develop an integrated set of fundamental probabilistic and statistical ideas which, among other things, will support informal inference. In this article, we describe the approach we have taken. We first outline four ideas that we claim lie at the heart of both data analysis and chance: model fit, distribution, signal-noise, and the Law of Large Numbers. We follow this with a brief description of classroom activities focused on *data analysis* that we have developed to highlight these ideas and their interconnections. But the major focus of the article is on how we explore these same ideas in explorations of *chance*.

We have recently finished a yearlong field test of these activities, working with grade 7 and 8 students at a local public middle school. We have just begun an analysis of the data from that field test, data which includes classroom videotapes, performance on items administered throughout the field test, and post-instruction interviews with individual students. In future articles we will explore those data in depth. In this article, we describe the approach and activities we have developed and our rationale for them. While we do not provide extensive examples of student responses or analyses of their thinking, we do offer some brief summaries. These are based on overall impressions recorded in reflective notes we made following class sessions. These impressions, however, should be regarded skeptically; our sense of what students understood as a result of the activities will certainly change once we have systematically

analyzed data collected during the field test. And the design of our activities will likely undergo revisions as we look more in depth at students' reasoning during and after the intervention.

2. IDEAS IN DATA AND CHANCE

We conceive of the activities we are developing as comprising two general strands — those that explore: 1) the "data in chance" and 2) the "chance in data." Four main ideas, which we view as lying in the intersection of data and chance, form the backbone of our activities: model fit, distribution, signal-noise, and the Law of Large Numbers. Below we briefly describe what we mean by these and summarize some of what we know from research about students' understandings of them. Although we describe these as separate ideas, they are highly interrelated. They are featured centrally (along with several other constructs) in Wild's (2006) succinct description of the statistician's view of the concept of distribution.

2.1 Four Main Ideas

Model Fit. In analyzing data, we have or develop expectations about features of the data. If we believe that a die is truly symmetric, then we expect to get about an equal number of each of the six sides when we roll it repeatedly. If we examine the heights of a sample of 16 year-olds, we expect the boys to be taller on average than the girls. When we look at data, we evaluate them with regard to our "model" (e.g., a guess, expectation, or prediction), sometimes rejecting the model based on the data. This is what we think Pfannkuch and Wild (2004, p. 22) point to when they emphasize the importance in statistical reasoning of the "recognition of the need for data."

Several researchers (Hancock, et al. 1992; Konold 1989; Watson, Collis, Callingham & Moritz 1995) have noted that some students occasionally ignore data in arriving at conclusions. But numerous studies report more encouraging results — that the majority of even fairly young students use data in sensible ways to evaluate prior predictions and expectations. For example, in Petrosino, Lehrer and Schauble (2003), nine-year-old students used data they collected as evidence that rockets with rounded nose cones reached higher altitudes than those with pointed nose cones, despite the fact that initially the students expected just the opposite. Pratt (2000) and Tarr, Lee, and Rider (2006) described 10- to 12-year-olds as systematically collecting data from

chance devices to decide on their fairness. In many cases, the results led the students to revise their initial judgments that the devices were fair. Thus, the idea of putting conjectures to the test by looking at relevant data seems to make sense to even young students (see also Masnick, Klahr, & Morris 2007).

Distribution. A collection of data (whether the sums obtained from rolling two dice or the heights of adult males) has properties that are different than the properties of the individual cases that make it up. If we organize numeric data as a frequency distribution, we can describe the distribution's shape, center and spread — characteristics that are not shared with an individual case.

Although the study of statistics concerns primarily these emergent, aggregate characteristics, numerous studies have demonstrated that novices tend not to focus on these features when interpreting data graphs (Ben-Zvi & Arcavi 2001; Hancock, et al. 1992; Konold, Higgins, Russell, & Khalil 2004). This research has inspired interventions targeting the development of aggregate reasoning in younger students. Those that have appeared successful (e.g., Bakker & Gravemeijer 2004; Cobb 1999; Lehrer & Schauble 2000) have these general characteristics in common. They:

- are rich in student discussion of qualitative aspects of distribution shape, including identifications of the location of hills or bumps, gaps, and spread-out-ness;
- keep students grounded in the real-world context from which the data came, so that when students begin noticing and talking about distribution features, the features are not disembodied characteristics of markings on paper, but are about the real-world situation under investigation;
- have students invent and compare alternative methods of data display and summarization and ultimately justify why they prefer to use one rather than another.

In short, effective interventions establish data collection and analysis as a modeling enterprise in which we need to evaluate choices in terms of the purposes they serve.

Signal-Noise. Separate samples obtained from rolling a die or from measuring the heights of a collection of females are unlikely to be identical. Because of chance variability inherent in the rolling of a die or the selection of particular individuals to measure, details vary from sample to sample. On the other hand, because a die is symmetric, the six sides will tend to come up equally often. And if we randomly sample from a population of females such that each individual has an equal chance of being in the sample, then the relative frequency of heights in a sample should tend to resemble the relative frequency in the population. Thus we can consider a sample as having two components — a "signal" that reflects the unchanging aspects of the population or process, and "noise" introduced via chance variability.

Konold and Pollatsek (2002) cited research reporting students' difficulties in using averages to compare groups as evidence that students were not regarding averages as signals. Based on historical analyses of the idea (e.g., by Stigler 1986; Porter 1986), Konold and Pollatsek (2002) suggested that repeated measures contexts, such as repeatedly measuring the length of some object, might provide a more suitable entry point for students into statistical measures. This is because in such contexts it is possible to associate signal-and-noise components of data to observable features of the measurement process. Several teaching experiments by Lehrer and his colleagues (e.g., Petrosino, et al. 2003) have made use of repeated measures and shown that students as young as ten can build up statistically sophisticated understandings of repeated measurements that are grounded on the conception that the measurements are composed of signal-and-noise components.

Law of Large Numbers. According to this law, as a sample gets larger, its aggregate characteristics converge on the corresponding characteristics of the parent population or process. Thus, the relative frequency of the various sums of two dice and the mean height of a random sample of adult males both settle down on their expected values as the sample gets larger, and thus provide better estimates than smaller samples of the actual population values. An important corollary of this law is that results that deviate substantially from the expected values are more likely with smaller samples than with larger ones.

Early research by Piaget and Inhelder (1975) suggested that by age 12, children expect that physically-created chance distributions (such as those generated using a Galton Board) will be more regular as the number of objects dropped through the device gets larger. This finding stands in contrast to studies by Tversky and Kahneman that indicated that adults usually ignore sample size when estimating the likelihood of events for which sample size affects the probability. For example, Kahneman and Tversky (1972) found that a majority of people believed that a large hospital is as likely as a small hospital to observe days on which 60% or more of the births are males (the smaller hospital is more likely to observe such days). However, subsequent studies (e.g., Sedlmeier 2007; Sedlmeier & Gigerenzer 1997; Well, Pollatsek, & Boyce 1990) have indicated that when posed differently, people can correctly answer questions such as the hospital problem. And several recent teaching experiments using technology suggest that the idea that distributional features (such as shape) of random processes settle down as the sample gets larger is indeed within the grasp of fairly young students (Pratt 2000; Stohl & Tarr 2002).

In developing these four ideas, the computer offers an especially powerful resource, both as a source of large data sets and also as a flexible and dynamic representational medium. Konold and Lehrer (2008) argue that computers offer dynamic forms of mathematical representation that spawn new avenues of thought. These dynamic representations also make accessible to even young students ideas and domains of explorations that, to this point, have been a high reach. Accordingly, as part of our project we have also been developing a probability simulation component that will be part of a future version of the data-analysis software *TinkerPlots* (Konold & Miller 2004). By housing probability simulation within a data-analysis tool, we provide a single environment within which students can explore the fundamental connections between data and chance.

2.2 Explorations of the Chance in Data

We have developed three types of activities for middle school students that highlight the role of chance in producing characteristic features in the distributions of data from 1) repeated measures, 2) production processes, and 3) different individuals (see Konold, Harradine, & Kazak 2007; Konold, Kazak, Lehrer, & Kim 2007; Konold & Lehrer 2008). Here we describe only one of these activities, which involves students in the exploration of repeated measures data.

In collaboration with Richard Lehrer and his colleagues, we have been developing activities in which students repeatedly measure some unknown quantity to try to determine its value (Lehrer, Kim, & Konold 2006; Konold & Lehrer 2008). In one such activity, we give students a footprint copied on a piece of paper, the print supposedly left at the scene of a crime. In their role as crime-scene investigators, their first job is to determine the length of the print. After they have performed various measurements, we show them a distribution (see Figure 1) that includes their measures along with data collected from other students. Their task is to come up with an estimate of the actual print length. With these particular data, most students agree on the value of 23.9, presumably because of the near perfect symmetry of the data values. We also discuss why not every measurer gets the same value. Having observed one another measuring, students can pinpoint several sources of error, including the accuracy of placing the ruler's 0 point, the angle of the ruler during the heel-to-toe measurement, and rounding decisions.

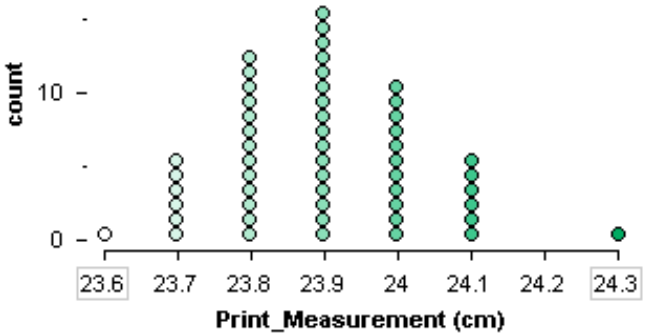


Figure 1: Student measures of a footprint of unknown length.

As we claimed earlier, we believe that situations such as this are ideal ones for introducing students to the idea of data as comprising signal and noise. Because the actual length of the print is not changing, it makes sense to students to attribute all of the variability in the measures they make to error (noise), and to believe that the true foot length is somewhere in the part of the distribution where the measures are most densely clustered. The later belief hinges in part on their understanding that errors can be positive and negative. Their experience in actually producing the measures plays a crucial role in supporting this view, as most of them can describe situations that would produce errors of each type. Their observations of one another measuring also allows them to identify and consider different factors that contribute to measurement error and to suggest ways to reduce the magnitude of errors. In this way, they can relate the amount of variability (the spread) to

different measurement protocols and thus expect more accurate measurement methods and tools to have less spread (Petrosino et al. 2003).

Building on these initial understandings, Lehrer, Kim, and Schauble (2007) have involved students in developing statistical measures of variation and in designing and critiquing simulation models of the repeated measurements built in the new development version of *TinkerPlots*. We have been classroom testing a related approach that involves showing students the Sampler shown in Figure 2 and asking them to predict features of the data (measurements) it will produce.

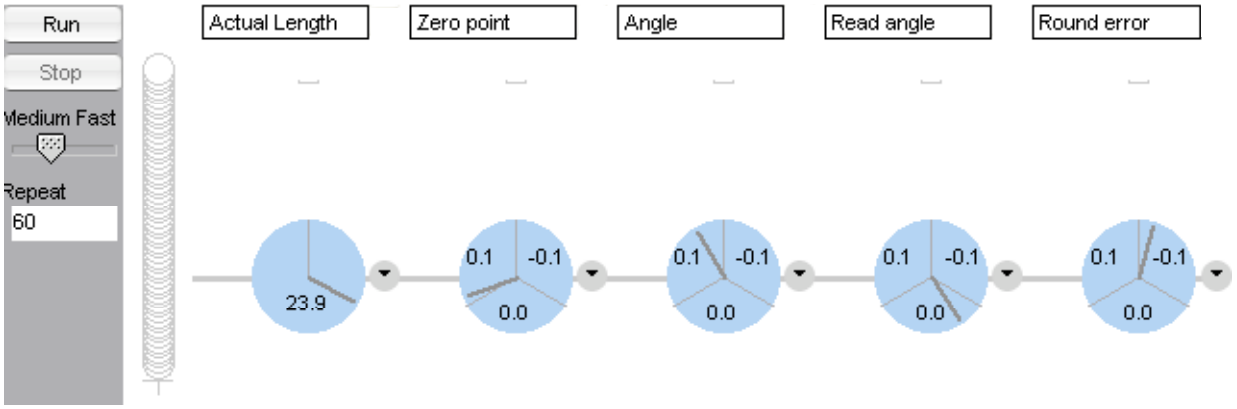


Figure 2: A *TinkerPlots* Sampler built to model 60 repeated measurements of a footprint. The first spinner gives the value 23.9 to a case. This is the estimate of the true length of the footprint. (Having only one element in a sampling device is currently the only way in the Sampler to assign a constant.) The other four spinners assign an error of 0, .1, or -.1 to each of four sources of error named by the students. Summing the values from these five spinners produces a simulated measurement value that is a combination of the true length and the measurement errors.

The spinner on the far left of Figure 2 contains an estimate derived from student measurements of the actual length of the footprint. (Though perhaps not ideal, currently the only way to generate a constant using Sampler is by including a single element in a sampling device.) The other four spinners, each representing a source of error the students have identified, assign an error value for that factor of either 0, -.1 or .1. After explaining the Sampler, we produce one measurement from it to demonstrate how a simulated measurement is generated in steps and then composed as the sum of the values from the five spinners. So the first run might result in the values 23.9, .1, -.1, 0, -.1, for a sum of 23.8. Once students understand the process, we tell them

that we are going to use this Sampler to produce 60 measurements and ask them to anticipate what the distribution of those 60 measurements will look like. Among the questions we explore is whether data from the Sampler displayed as a frequency graph will resemble in shape the distribution of real measurements (Figure 1). Many believe that while the Sampler will produce measures of roughly the same range, that the distribution will be relatively flat and not peaked at 23.9. They are therefore initially surprised to see that the Sampler gives them a shape similar to the one they observed for their actual measurements. To see whether this shape persists, we repeatedly run the model and look at several examples (Figure 3). After-the-fact, many students can offer a reasonable explanation for the shape — that it is rather hard to get extreme measures, because you have to be very "lucky" or "unlucky" to get the same error value (all -.1s or all +.1s) from the four error spinners. Combinations of the three possible values, which give sums near 23.9, are more likely.

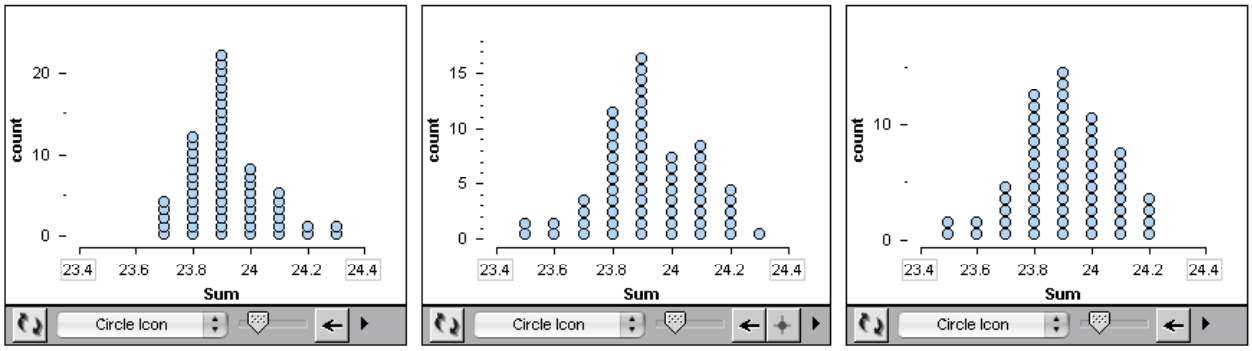


Figure 3: Three samples of 60 measurements generated from the Sampler in Figure 2. The values of Sum (the modeled measurements) were determined by summing the values of each of the 5 spinners for each of the 60 repetitions.

At this point in the activity, there are two instantiations of signal and noise. One is the single distribution in which the signal is expressed as an average, and noise corresponds to the distribution of values around that signal. But we can also regard the *shape* of the distribution as a signal, caused by the independent contributions of error factors. The noise in this case is the variation in the shapes from one sample to the next. While each of the samples is unique, we can also begin to see a commonality of shape.

To address the question of whether the foot measure Sampler is a good model of actual measurements, we need a basis of comparison — "good" compared to what? To help move the students in this direction, we invite them to modify the Sampler so that it will make data that differ in noticeable ways from their actual data and/or to anticipate how various modifications will affect the output. It is relatively easy for them to produce distributions that are centered on different values and to increase the range. Making alterations that change the basic shape are more challenging and take students in the direction of considering bias in measurement. Thinking about how to reduce variability while leaving the number of error factors at four helps them think about the implications of fine-tuning a measurement process. With a collection of bad models as a basis of comparison, students can now engage in a meaningful conversation about whether and why the Sampler in Figure 2 is a good model and can suggest ways in which it might be improved.

To summarize, our hope is to support students viewing repeated measurement data as a combination of signal and noise and to help foster a basic understanding of why distributions of such measurements are centered on, and clustered around, the true value (see Wilensky 1997).

2.3 Explorations of the Data in Chance

In the remainder of the article, we describe our approach to introducing students more formally to ideas of chance and probability while trying to maintain a focus on ideas that relate to and support data analysis. Probability comes with its own particular learning challenges, which we briefly consider as a backdrop for our approach.

It is common to encounter in introductions of texts on probability an acknowledgment that, to many, the topic embodies a contradiction. How can we say that coin flipping is an inherently unpredictable, chance process and then turn around and assert, based on mathematical computations, that the probability of flipping heads four times in a row with a fair coin is 0.0625? Precise numeric statements can seem decidedly out of place in the messy domain of chance.

In the domain of mathematical modeling, discomfort of this sort comes with the territory. In her book analyzing the history of the early development of probability, Daston (1988) observed that "Recasting ideas in mathematical form is a selective and not always faithful act of translation" (p. 4-5). Probability theorists of the late 17th and early 18th centuries were, according to her, guided initially by the degree to which their formalisms captured and modeled the world as they understood it. She examined how early probabilists "bent and hammered their definitions and postulates to fit the contours of the designated phenomena with unusual care" (p. 5). The ultimate fate of the formalisms hinged on how well they predicted the results of, and explained, the phenomena they were designed to model. But that was over 300 years ago, plenty of time to forget that it was only after years of considerable effort that probability could be perceived by its adherents as "common sense reduced to calculus" (Laplace 1814, as quoted by Daston 1988, p. 68).

Probability curricula trade in such historical amnesia. As a result, they rarely leave room for students to acknowledge and resolve thorny issues that arise as they are instructed in how to apply ready-made mathematical tools to carefully-carved pieces of the world. Steinbring (1991) suggested that a weakness of traditional probability instruction is that it presumes that by offering students clear definitions of probability from the beginning, a sound foundation can be established on which basic probability concepts can then be developed. Steinbring based his critique of this practice on the observation that probabilistic interpretations evolved historically in relation to one another and in response to disjunctions between objective and subjective interpretations (cf. Hacking 1975). He maintained that students' understandings also develop in a dynamic, non-linear way despite instructional attempts to eliminate ambiguity and false steps. Accordingly, Steinbring advocated an instructional approach in which ambiguities and seeming paradoxes are explored rather than finessed with the hope that from such fluid explorations, coherent and robust probabilistic understandings can emerge.

While agreeing in principle with Steinbring's (1991) analysis, the pedagogical vision he offers is difficult to imagine working in classrooms where most teachers are, themselves, trying to come to grips with the domain. Our approach takes a different tack to address Steinbring's concerns. We proceed by first establishing a set of shared observations that students, as a group, work to

understand, offering tools and formalisms only after students perceive the need for them. Thus, we do not start by providing definitions of events, sample spaces, and probabilities, but by having students collect data that reveal an unexpected pattern of results, results that beg for an explanation. In the process of further exploring the phenomenon, we introduce students to techniques and formalisms as possible explanatory tools, but leave it to them to test, discuss, and ultimately judge their usefulness. In this respect, the approach is similar to Daston's description of the approach taken by early probabilists in which formalisms were crafted and adapted to fit the field of application, rather than the other way around. Our instructional approach is highly dependent on the computer's power to generate large samples of data quickly, both to establish the phenomena that beg to be explained and also to test the ability of probability models to explain them.

Our approach, evolved through several iterations of classroom testing, incorporates several of the suggestions offered by Scheaffer, Watkins, and Landwehr (1998), chief among which was the exhortation that "the unifying thread throughout the probability curriculum should be the idea of distribution" (p. 17). The focus on distribution is consistent with Shaughnessy's (1997) exhortation that we spend more instructional time focusing on variability and less on center. It is exemplified in probability instruction described by Abrahamson (2006) and by Hovarth and Lehrer (1998). In their classrooms, students do not compute probabilities of events but rather explore the shapes of distributions (such as the sum of two dice) obtained from repeated trials where a major objective is for students to come to understand these shapes in terms of the number of ways the various events can occur (i.e., the sample space). There are many advantages of focusing on distribution features rather than event probabilities. First, considerable research has established that students tend not to focus on aggregate features of distributions, but rather on the values of particular cases or the frequencies of selected outcomes (see Konold, et al. 2004). We hope that by focusing on distributions and their shapes we will prepare students to attend to distributional features when examining data. By focusing on general shape we also are able to finesse, for the time being, some common but erroneous beliefs that students have deduced from experience, such as the belief that a single die is less likely to come up a 6 than a 1 (see Watson & Moritz 2003). Second, focusing on distribution shape facilitates students eventually seeing the relationship between the data they generate and the sample space they construct. Consider what

is involved in evaluating the probability of a particular event. To determine the relation between the sample space and an actual result obtained in a sample requires computing and comparing a ratio of a subset of the sample space (say 6/36 for the sum of 7 of two dice), to the ratio of 7s in a particular sample (say 18/100). To do this requires good facility with rational numbers, which many of the middle school students we work with do not yet have. On the other hand, and as we describe later in the article, students can relatively quickly assess the qualitative fit between the general shape of the expected distribution based on the sample space, and the shape of the distribution of actual results. But perhaps the most important reason for focusing on distribution shape rather than event probabilities is that distributions simultaneously communicate the variability in the outcomes of the situation and the structure that begins to show through this variability as we add more and more data to the distribution. Thus, the distribution emphasizes the variability inherent in a chance process, which seems to be the main idea students hold about chance prior to instruction and, thus, a fruitful place to begin.

3. OVERVIEW OF INSTRUCTION

3.1 Classrooms

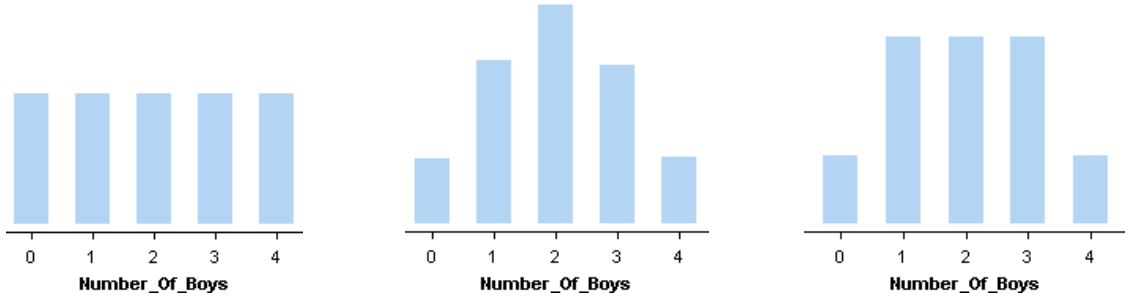
We have conducted four rounds of field tests in classrooms at Lynch Middle School in Holyoke, Massachusetts. The field tests from which we draw most heavily in this article were conducted during the 2007-08 school year. During this period, we co-taught three 10-week classes devoted to the topics of probability and data analysis. Each class met every other day in 70-minute sessions in a classroom outfitted with laptop computers and a computer projection system. The classes included an 8^{th} grade class (11 students) and two 7^{th} grade classes (one with 8 students, the other with 9). Lynch Middle School serves about 400 students who are predominately minority (76%), with 85% of the students qualifying for free and reduced-priced lunches. Schools in the Holyoke School District have some of the lowest scores in Massachusetts on the high-stakes MCAS test.

3.2 Three Probability Contexts

During the ten-weeks of instruction, we interwove activities dealing with data with ones dealing with chance. During this time, we involved the students in three, separate, multi-day chance activities, briefly described below.

The Wink Game involves blindly drawing twice, with replacement, from a bag that contains two disks. One disk is labeled with a dot (•), the other with a dash (-). If both draws result in a dot (•,•), Player A wins. If both draws are dashes (-,-), Player B wins. Player C wins if the two draws are different (•,- or -,•). Imagining the symbols to be the two eyes of a person's face, we refer to the mixed outcomes (•,- and -,•) as "Wink," to two dashes (-,-) as "Blink," and to two dots (•,•) as "Stare." In addition to injecting a bit of humor into the game, using these names allows us to refer to the events in a way that provides no hint as to the critical role of the order in which the single outcomes, dash and dot, occur. The question we pose to students before playing the game is whether the game, with three players, is fair. Initially, most students believe it is.

In the Family Problem, we explore the composition of the genders of children in families with exactly four children. After we examine a few specific families and place them on a distribution according to the number of boys in the family, we ask students: "If we graphed the number of boys in many, many real families with four children, which graph below do you think would look most like the real data?" (see Figure 4).



Graph 1 Graph 2 Graph 3

Figure 4: Three possible distributions for the number of families with 0-4 boys. Many students initially select Graph 3. A common argument is that it is very hard to get either all boys or all girls, but the other three results should be about equal.

In the Dice Problem, we ask students to select from the alternatives shown in Figure 5 the distribution we would most likely get if we rolled two dice 1000 times and plotted the sums, 2-12.

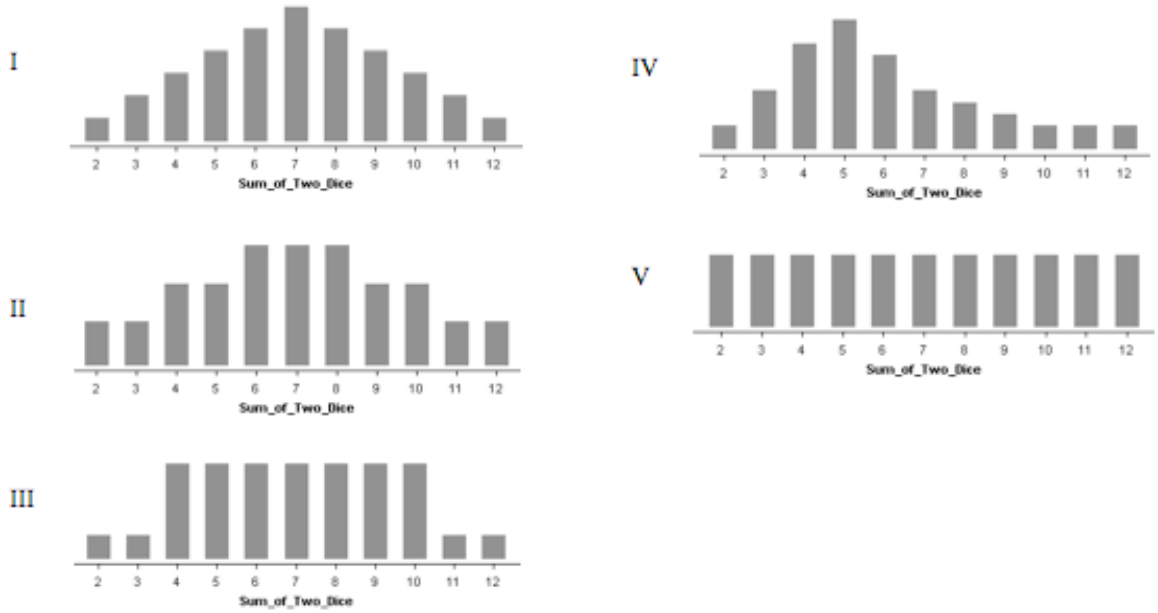


Figure 5: Five possible distributions for the sum of two dice. Most students initially select graph III. A common explanation is that sums at the extremes will not occur very often but that the other sums will occur about equally often.

There are a number of criteria we used in choosing these particular problems. All three contexts are approachable both through theoretical analysis via the sample space and through conducting physical trials. This allows us to explore the relationship between the two approaches. Accordingly, we have not spent much time having students explore chance set-ups that involve non-symmetric elementary outcomes, such as the tossing of thumbtacks (e.g., Newman, Obremski, & Scheaffer 1987, p. 8). We agree with Scheaffer, Watkins, and Landwehr (1998) that the study of counting methods has little to do with understanding probability. Thus we have restricted ourselves to situations where the sample space is relatively small such that students are capable of generating all possibilities by simply listing them.

To facilitate the application of the idea of signal-noise, we chose contexts and the particular questions to pursue so that even in relatively small samples, a trend is already visible. All three investigations pose a question about compound events whose distribution is non-uniform. It is

the general shape of the distribution that surprises students and thus motivates further exploration and explanation. We have not had students investigate uniform distributions, such as the question of whether if you roll a fair die the results will be about equal across the six outcomes (e.g., see Tarr, Lee & Rider 2006). Deciding whether simulated or real data provide support for a uniform distribution is a difficult assessment for young students (and even adults), as some outcomes will almost always appear to them to be occurring more than expected and others less than expected (see Figure 6). This makes it hard to perceive the signal (the distribution's on-average, uniform shape), through the noise.

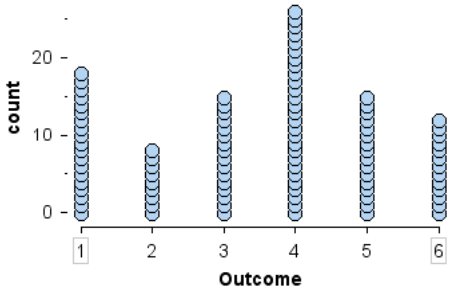


Figure 6: Simulated results of 100 rolls of a "fair" die. Students asked whether such data suggest that the outcomes are occurring about equally often will invariably conclude that some outcomes appear more likely than others. Students (and many experts) do not appreciate how much variability there is in random samples of this size. The "signal" in non-uniform distributions is much easier to perceive informally even at small sample sizes than are the signals in uniform distributions.

4. CHARACTERISTICS OF OUR APPROACH TO PROBABILITY

All three of the investigations we have designed target the four central ideas we previously described — model-fit, distribution, signal-noise, and the Law of Large Numbers. In this section we outline general characteristics of each investigation that support student learning, citing examples from all three contexts. In brief, each investigation begins by establishing the need to explore the situation by keeping track of the results of repeated trials. These data generally are inconsistent with students' initial expectations. To help explain the pattern of results they get, students build an "expected" distribution by generating elements in the sample space, arranging them as a distribution, and then using the computer to see how well this "expected distribution" fits the distribution of simulated results. The investigations conclude with students comparing the relative fit of the expected distribution to data from samples of different sizes.

4.1 Making and Testing Initial Predictions

We begin each probability investigation by having students make guesses about the situation. These guesses are based on their current understandings (or theories), which include primary

intuitions (Fischbein 1975) such as the equiprobability bias (Lecoutre 1992), the representativeness heuristic (Kahneman & Tversky 1972), and the outcome approach (Konold 1989). For example, after describing the Wink Game and selecting three students to play it in front of the class, we ask students to decide whether the game is fair and to write a brief explanation. Many students initially regard the game as fair based on the fact that there are three possibilities and that chips are blindly drawn from the bag after mixing them well. To introduce the Family Problem, we collect data by asking students to write on index cards the composition of boys and girls in four-child families they know. We then add their cards to a stack of cards we have created beforehand to give a total of about 50 cases (see top of Figure 7). Before looking at the distribution of real data, students make their predictions about the distribution of the number of boys in four-child families by choosing one of three sketches (see Figure 4). After arranging the 50 cards according to the number of boys along a number line taped to the floor (see bottom of Figure 7), students discuss and revise their predictions based on the data. This distribution is different enough from what most expect that they begin to question their initial conjecture. But at this point, they have no clear explanation for why the distribution is shaped the way it is.

Figure 7: Students record on individual cards the gender orders of families they know of with four children (top). They arrange these to form a frequency distribution of the number of boys in a family (bottom).

There are several reasons for having students make initial predictions. First, it provides additional motivation for exploring the problem. Second, having explicit expectations prepares students to observe specific features of the data that otherwise might go unnoticed (see Konold 1995). Most importantly, predictions serve to establish the purpose of the activity as understanding the situation and set in motion the activity of model fitting.

On each problem, after students make their predictions, we manually collect data from the process (e.g., we roll real dice). This is a critical phase because it helps to ensure that the students understand the process before they go on to model it. For the Wink Game, after playing the game once in front of the class, students break into pairs. Using a bag and two chips, each pair plays the game 12 times and records their results. We have found that with our problems (all of which involve a repeatable chance experiment), students are quick to evaluate their initial judgments (correct or not) on the basis of data they themselves collect. Indeed, as others have reported (e.g., Aspinwall & Tarr 2001), many of the students express confidence about conclusions drawn from quite small samples. Accordingly, we delay formal discussion of data until the sample under consideration is relatively large.

In the case of the Wink Game, we accomplish this by recording the data collected by each pair on the classroom whiteboard (see Figure 8). In this table, each of the five columns contains the results of the 12 trials conducted by a pair of students. The sixth column, added at the suggestion of the students, shows the total wins for each event across the five groups. When asked what they conclude from their combined results, most students will now assert that the game is not fair — that Wink has a better chance. But they will also acknowledge that, due to the variability evident in the results of each pair, more data would help them decide with confidence whether the game is really fair or not. Thus, the sample-to-sample variability they see in the results of different groups seems to temper the confidence they would express if looking only at their own results.

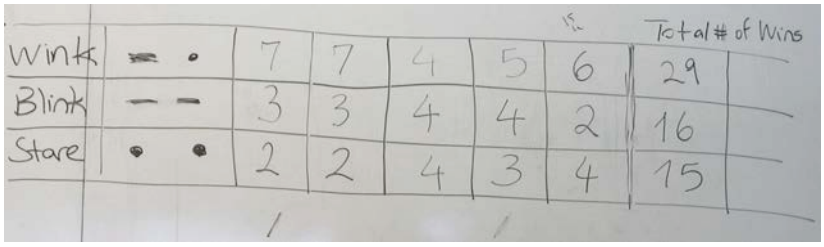


Figure 8: Table showing the results of five groups each of which played the Wink game 12 times. The rightmost column shows the combined totals for each event.

We have replicated the Wink Game activity in six different grade 6-8 classrooms, and the student responses have been quite consistent. Thus it seems, given this context at least, that even before instruction students have a general sense of model fit that helps them anticipate and interpret variable data. That is, they can anticipate the sort of data that will support their initial conjecture (in this case, that the three events will win about equally often) and will question that conjecture

when the data are sufficiently different from what they expect. Furthermore, when confronted with the results of different groups, they are not surprised to see that the results vary and they try to perceive a general trend in them. Konold (1989) reported that some university students questioned the usefulness of data in deciding which side of an irregular-shaped bone was most likely to land upright. In particular, they did not make use of frequency data from 1000 rolls since the results of another 1000 trials would be different. By contrast, in the context of these three activities, students make use of data in deciding whether their initial thinking provides a good fit with the data. No student has argued that the variability in the results rendered the data useless. This also suggests that students have a basic notion of signal and noise even before instruction in that they can evaluate a particular sample in terms of its general trend and will regard deviations around what they expect as the contribution of chance.

4.2 Testing Predictions with Simulated Data

To further test their predictions, students use the development version of *TinkerPlots* to build a model of the situation that they use to quickly generate and analyze large samples. We first introduce students to *TinkerPlots'* simulation component (called the "Sampler") by using it to randomly select three students in the class to play the Wink Game. To do this, we create a mixer, as shown in Figure 9, that contains each of their first names. Working together as a class, we edit the mixer contents so that only students present that day are in the mixer and set it up so that three names will be drawn without replacement.

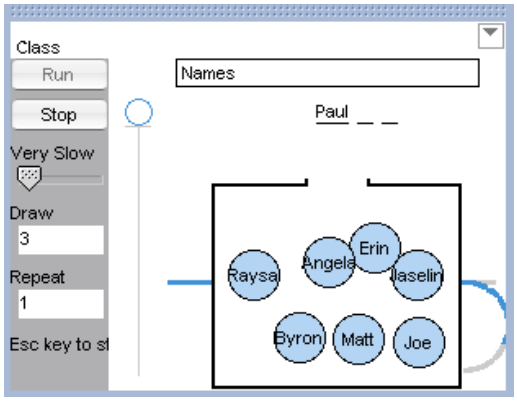


Figure 9: To select three students to participate in the Wink Game, we fill a *TinkerPlots* Sampler (a mixer in this case) with balls labeled with the students' names. Pressing the Run button causes the balls to bounce around in the mixer. Three balls are randomly selected, one at a time. When a ball is selected, its name floats up into one of the three slots above the mixer. In this example, the first name chosen was "Paul." The second name is in the process of being selected. The top of the mixer is open to indicate that the sample of three is being drawn without replacement.

Thus we introduce the Sampler to students not to model a chance phenomenon but as a way to fairly choose students from the class. The idea that we use the Sampler to model a chance event is first suggested by the students. This happens in

the context of the Wink Game when we suggest to students that, to collect more data, each group play 100 repetitions of the game. They protest that it will take them too long, and finally one of the students will suggest using *TinkerPlots*. When we ask them how to do this, they instruct us to remove all but two elements, to label these with a dash and a dot, and to draw twice with replacement. We ask whether they are sure they will get the same sort of results from this Sampler as they do from the bag. Perhaps still worried that otherwise we will have them conduct 100 physical trials, they express their confidence that playing the game with the computer version is exactly the same as playing it with the bag filled with two chips.

Note that in this sequence of instructional activities, the students see the computer modeling as a logical next step, motivated by the need for more data and the time required to collect it by hand. Indeed, our aim in designing the sequence of activities in each of the investigations is that students see each subsequent phase as logically following from the previous one, providing a next sensible step in the exploration.

In building a model of the situation in *TinkerPlots,* students work individually at computers. For the three problem contexts, students generally have relatively few difficulties building an appropriate model. The software allows them to build Samplers using mixers, spinners, or distribution objects, and to draw either repeatedly from the same device, or once each from a sequence of in-line devices. Figures 10 and 11 show three models, one for the Wink Game, and two for the Dice Problem.

For the Wink Game, most students use one mixer, as it closely resembles the real situation where two discrete objects are drawn from a single bag (see Figure 10). In the case of rolling two dice, the two-spinner model (see right side of Figure 11) bears a closer resemblance to the real situation than does the one-spinner model (left side of Figure 11). As with real dice, the two-spinner model involves two objects where each object has six different "landing" positions. Students have different preferences in the Samplers they build. Thus one student will often be working next to another who uses a different device type and/or in-line vs. looped sampling. When questioned, however, they will generally say that the differences between their models are not consequential — that both Samplers will generate the same sort of data.

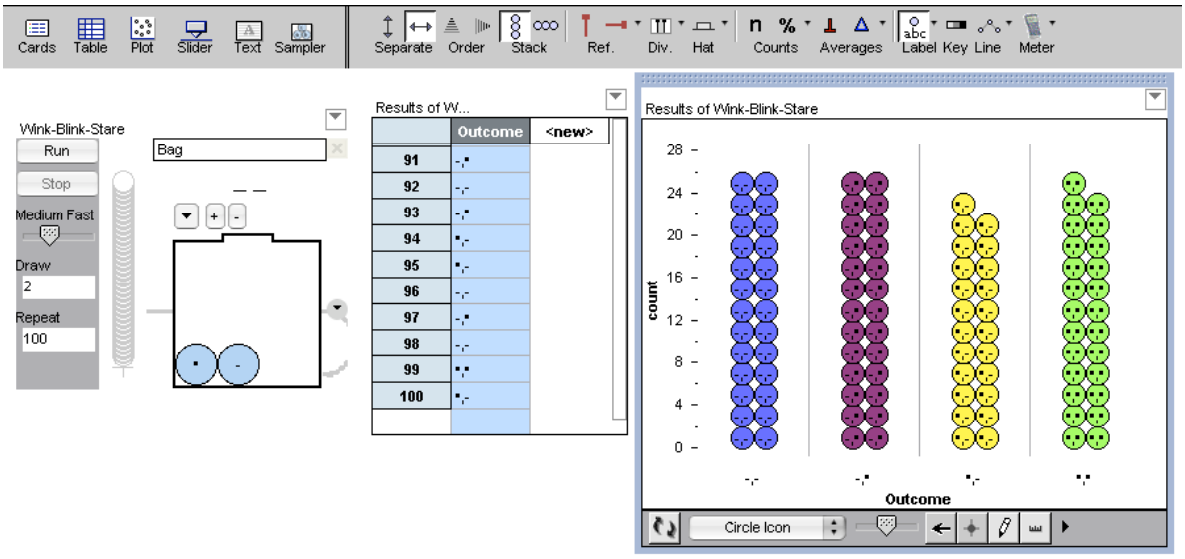
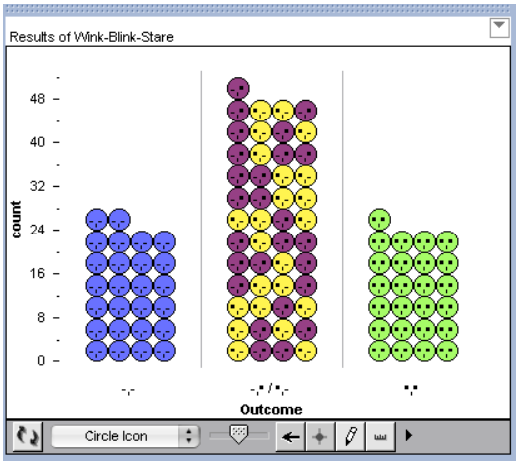


Figure 10: Computer model of the Wink Game. A single mixer (upper left) is set to draw twice from the mixer a total of 100 times. The results table to the right of the mixer displays the repetitions as they occur. The plot at the upper right shows the number of outcomes of the four types in that sample of 100. In the graph at the bottom, the outcomes -,• and •,- have been combined into one bin by dragging one into the other. Surprisingly, students creating and interpreting these graphs do not in the process come to the realization that the reason Wink occurs about twice as often as Blink or Stare is that it comprises two different outcomes.



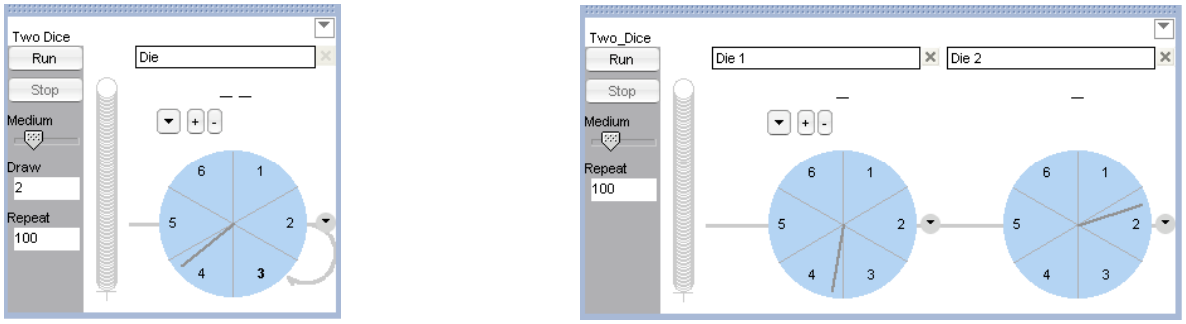


Figure 11: Two alternative methods of building a *TinkerPlots* Sampler to model the rolling of two fair dice. In the Sampler on the left, a single spinner spins twice (draw number = 2) to execute a single trial. The loop at the right of the spinner indicates that the spinner will spin more than once per trial. The Sampler on the right uses two spinners, each spinning once to execute a trial of two rolls. The grey column at the left of the spinners is a stack of 100 balls (trials). As the sample of 100 is drawn, the stack gets smaller.

Critical to the success of our instructional approach is that the students view the computer models they build and run as generating the same sort of data they get when they perform experiments with the actual physical devices. We designed the animations in the Sampler to support the belief (a fiction, of course) that the computer devices are working fundamentally the same as the corresponding physical devices: balls in the mixer bounce around randomly and with realistic motion; the arrow on the spinner rotates quickly at least one revolution and then gradually slows and stops. We suspect that most of the students know that the choosing is happening deeper in the computer code. Nevertheless, students judged our early animations as being nonrealistic and argued, therefore, that the samples drawn from them were not like the samples they would get from, e.g., real dice. So it seems that realistic animations are critical in supporting a trust that the computer is indeed drawing randomly. We occasionally question students' beliefs about the computer-generated data, and most of them seem to accept their veracity.

We also believe that critical to the acceptance of the computer models is the students' initial experiences with the real phenomena (chips in a bag; real dice). When they do move to the computer, the fact that the data they get from it resembles the data they have already collected from the real situation undoubtedly lends support to their belief that the computer provides a veridical model.

Using the Sampler, students draw large samples which give them more confidence that their initial predictions are incorrect. While most will say that, for example, Wink is twice as likely as Blink or Stare, they are not generally able at this point to offer a compelling argument as to why this is. This sets the stage for an exploration of the sample space.

4.3 Using the Sample Space to Explain Distribution Shape

Involving students in the development of the sample space serves several purposes in our activity sequence. First, it helps to explain the data students have previously collected either from the real situation and/or from the computer simulation. In the Wink Game in particular, the sample space provides an explanation that satisfies students as to why, contrary to their initial predictions, Wink wins about half the time. But as important, the sample space, organized as a distribution of

event types, prepares the students to perceive data they subsequently get from the Sampler in a more statistically structured way — as a noisy version of the distribution they expect (the signal) based on the sample space.

To develop the sample space for the Family Problem, we begin by giving the students a task inspired by Abrahamson (2006). We hand out a supply of green and blue stickers along with strips of paper composed of a row of four squares. The students' task is to create as many different "4-bars" as they can by placing stickers in the blank spaces. To help them keep track of configurations already made, about half of the students will begin organizing the completed bars by type: those with 1 blue-3 green, 2 blue-2 green, etc. Thus in the end, many of them have created a distribution that orders and organizes the 16 possible bars by type (see Figure 12). After generating this representation, students can quickly appropriate it to thinking about the number of possible four-children families, the colors green (G) and blue (B) facilitating this reinterpretation.

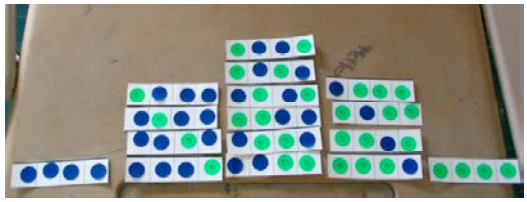


Figure 12: The 16 possible four-bars generated by a pair of students and organized by type (i.e., by the number of blue stickers). During the construction process, this emerging organization helps the students to keep track of bars already made and to spot omissions.

In the case of the Wink Game, as soon as we ask students to list all possible results, about half of them will immediately list the four simple outcomes. Those who maintain that there are only three possibilities are ultimately convinced by an argument one of the students will typically come up with: as soon as the first chip is drawn from the bag, either the student with Blink or the student with Stare is eliminated. But the student with Wink is always still in the game after the first draw. This also leads them to the insight that Wink will win about half the time. Once it has been made explicit, students quickly see the sample space as explaining the results they have collected, saying things such as "Wink occurs twice as often because there are two ways to get it."

After constructing the sample space of the Wink Game and seeing that there are four possible outcomes, we use a model built in the Sampler to generate more data in front of the class. The

data we generate are similar to the data they have each collected from the models they built and ran. We first display the data as shown at the top of Figure 13. We then color the cases by the attribute "Outcome," (graph b) and ask them what they notice about the central, "Wink" column. They note that it is made up of two different colors. We now order the cases to produce the graph in c, and finally separate Wink into its two constituent classes (graph d). Our purpose is to support the students' perception that the Wink column of the data is composed of the two outcomes, which explains why it is about twice as high as Stare and Blink. This perception, we believe, is not quite the same as understanding that there are two ways, in theory, to get Wink. In these graphs students can see that actual data can be recomposed to show the four simple outcomes, two of them comprising the event "Wink."

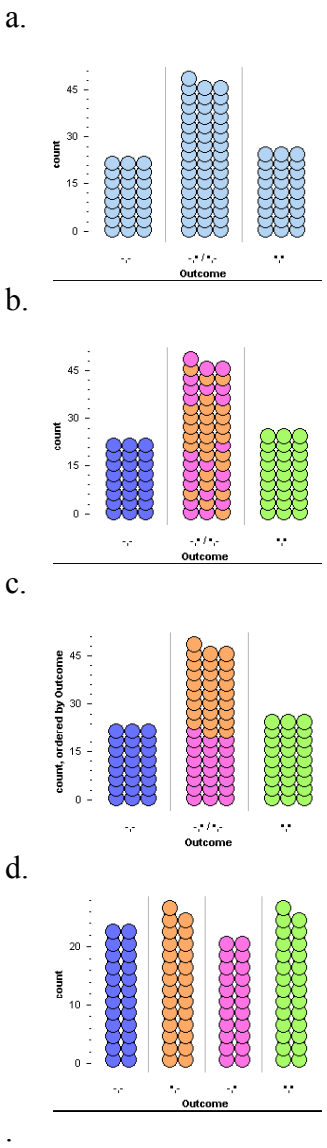


Figure 13: Results of 100 simulated trials of the Wink Game graphed in four different ways in *TinkerPlots*.

In our first classroom testing of the Wink Game, we expected that many students in producing their first graphs of their data in *TinkerPlots* would come to see that there are two ways to get Wink even before we formally explored the sample space. As you can see in Figure 13, *TinkerPlots* colors the two outcomes comprising Wink differently, and in making a graph showing the three winning events, students must physically drag the two outcomes together. Additionally, the name of the value on the graph is not "Wink," but is made up of the concatenated labels of the individual outcomes (e.g., •,-/-,•). Surprisingly, only 2 of the roughly 70 students we have worked with have, as a result of operating on the graphs, come on their own to perceive as different the two outcomes that comprise Wink. The other students combine the two outcomes in the graph

because “they are the same thing,” so perhaps it makes sense to them that there are no important differences between the two outcomes that they need to consider. Students make a comparable fusion when they are recording the data from their physical trials. In the worksheet we provide, they record the outcomes as they occur, preserving the order information. But labeling both outcomes “Wink” apparently supports the idea that the two outcomes are indistinguishable, or at least that the difference between them is inconsequential.

4.4 Seeing Signal and Noise in Empirical Distributions

Each of our investigations concludes by exploring an issue that generally gets raised early in each investigation. This is the observation that when individual students collect data, they each get something a bit different. At some level, this fits with their expectations — they understand that when chance is involved, you do not get the same thing every time. We aim to build on this intuition, however, to help them come to see in data a noisy version of their expectation.

To help support this view, we have found it necessary first to encourage a more holistic perception of data displayed in graphs. In our initial classroom tests, students would want to record exact frequencies for each event even in samples of 1000, and would report small differences between the heights of two bars as significant. To reorient their perceptions, we now instruct them in how to make quick “sketches” of data.

We introduce sketching as a way for students to quickly record the results of a simulation. Pointing out that it would take a lot of time to draw on their worksheets exactly the graph they got or to count and report actual frequencies, we demonstrate quick sketching on the board. We point out that in making a sketch of a distribution, we pay attention only to the overall shape and to relative heights of bars or stacks. Following this instruction, we use the computer to draw large samples quickly and give the students about five seconds to sketch the graphs. We also play a game where one student of a pair makes quick sketches of five different graphs and the other student, whose back has been turned, has then to match those sketches to their source graphs. The game not only gives them more practice making sketches but also demonstrates that even sketches made very quickly can capture salient features of a distribution that one can use to pick it out of a “line up.”

The culmination of each investigation involves students using the computer models they have built to draw multiple samples and compare the distribution of results they get to the distribution they expect based on the sample space. Figure 14 shows the worksheet students used for the Family Problem.

In this example, students draw a total of 4 different samples of 160 simulated families. For each sample, they make a quick sketch of their results and then rate the degree of fit between that distribution and the one they expected based on the sample space (shown on the left of Figure 14). With samples of size 160, most of the results, from students' perspectives, tend to produce "Great" and "OK" fits. As part of the subsequent classroom discussion, we pose the question of whether the fit will be better or worse if we looked at samples of size 32. A few think they will be worse, but more of them predict that the fit will be better, perhaps thinking (correctly) that the chance of getting a *perfect* match to the expected values is greater. Students seem surprised when we run these simulations and observe mostly examples of "Bad" fits (see Figure 15).

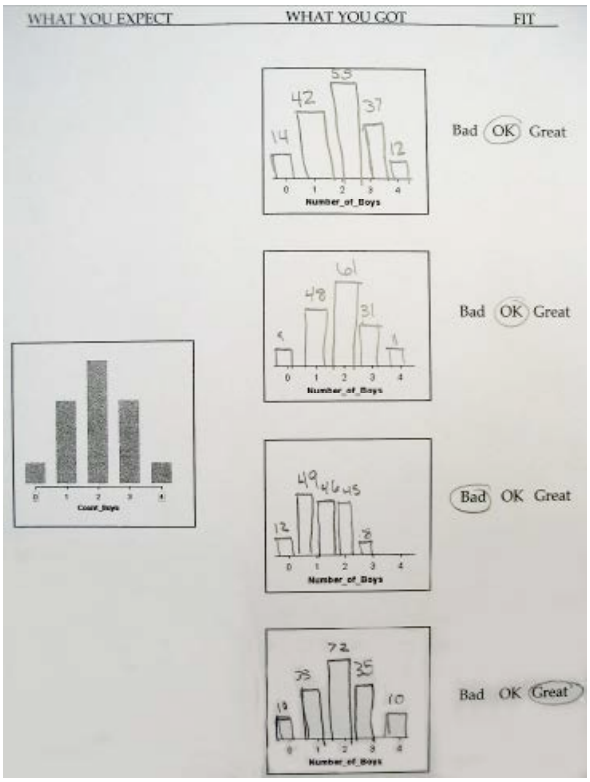


Figure 14: On the left is a graph made by organizing the 16 simple outcomes in the sample space for the Family Problem. The label "what you expect" suggests that students use it as an expectation for the shape of the distribution they will get when running the simulation. On the right are four empirical distributions, which a student sketched. These summarize results of four replications of a *TinkerPlots* simulation, each replication creating 160 "families." For each observed distribution, the student has rated the degree of fit to the expected distribution. Note that this student was not satisfied with sketching the results; she also recorded the actual numbers above each frequency bar. Weaning some students from attending to details is not easy.

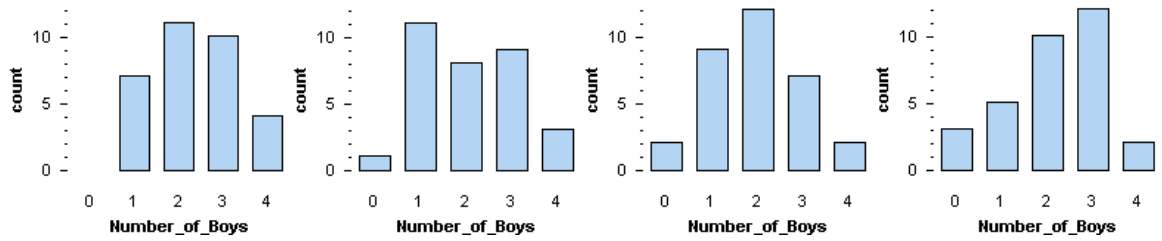


Figure 15: An example of the types of results obtained from running successive simulations of the Family Problem with small samples (here n=32).

At this point, students request that we run both smaller and larger samples to see what happens. When we ask about their expectations regarding larger samples (e.g., 2000), most do not anticipate how closely they will resemble the expected distribution. The close fits of the larger samples (see Figure 16) generate considerable discussion and the desire for an explanation.

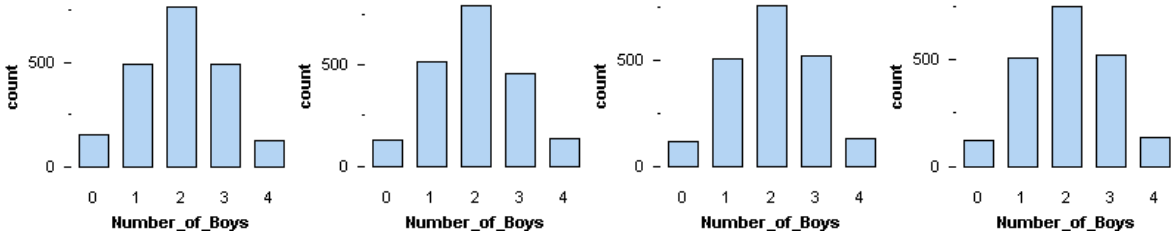


Figure 16: An example of the types of results obtained from running successive simulations of the Family Problem with very large samples (here n=2000).

Our perception of closeness hinges in part on the scaling that the graphs in *TinkerPlots* do automatically and on the fact that we are attending to global features revealed at the "sketch level" rather than to the absolute numerical differences among the event frequencies (see Hovarth & Lehrer 1998, p. 131, for a related discussion). Surely, most of our students do not appreciate the effects of scaling on their perceptions. We provide students a more qualitative explanation of the Law of Large Numbers, telling them that with large samples, we tend to get what we expect. With small samples, however, it is fairly easy, just by chance, to get results that deviate from what we expect. This is because just a few single outcomes in a small sample can radically affect the shape of the distribution. We follow this up by doing a classroom demonstration in which we repeatedly watch a sample of 2000 as it grows over time. The observation is that at the beginning, the distribution often looks little like the expected distribution but that, after time, it comes to closely resemble what we expect according to the

sample-space analysis. When we encounter this problem again in other contexts, many of the students can quickly anticipate the relationship between sample size and model fit.

4.5 Summary of Approach

In the yellow and blue portions of Figure 17, we depict our view of the structure of students' explorations of each of the three problem contexts. In this approach, we alternate repeatedly between phases of theory exploration and development (yellow portion of Figure 17) and data

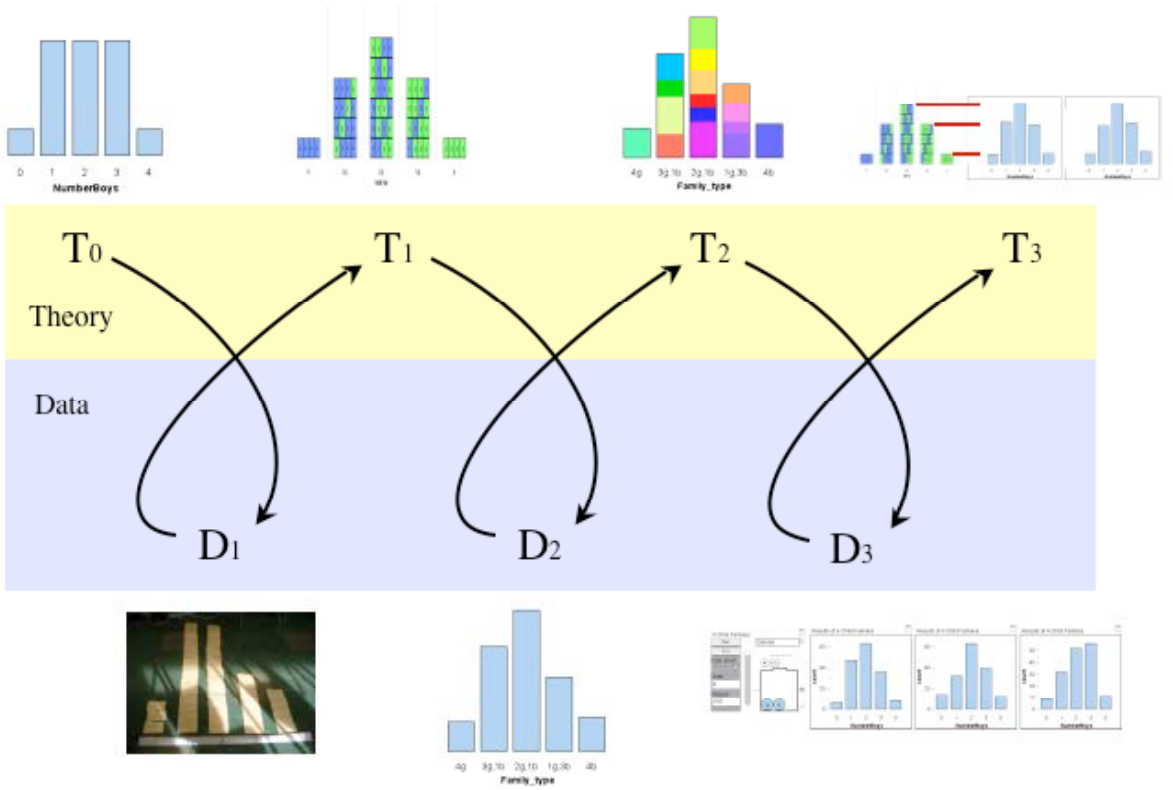


Figure 17: Theory-data phases in our chance activities. This iterative structure between theorization (upper part of diagram) and data (lower portion) begins with students testing initial predictions (T_0) by collecting data from the real situation (D_1). Theorization typically progresses with an exploration of the sample space (T_1), which students use both to explain the data they have previously observed and to formalize what they expect to observe in future samples, generated using *TinkerPlots* (D_2). Finally, students use their computer model to repeatedly test how closely simulated data (D_3) resemble their theoretical expectations. In the process, they come to understand sample-to-sample variation in terms of random noise (T_3), which they can manipulate via the sample size (i.e., the Law of Large Numbers).

collection and analysis (blue portion). Each investigation begins by having students make their best guess about the particular situation. For example, in the Family Problem, most students

initially expect the distribution shown above T_0 in the figure. This expectation might be based on a combined belief in the equiprobability bias (Lecoutre 1992) and the representativeness heuristic (Kahneman & Tversky 1972), according to which students regard the extreme values as unlikely but all other possibilities as equally likely.

After they have made their expectations explicit, students then collect and analyze data from the real situation. In the Family Problem, this results in a distribution like that shown at the bottom left of Figure 17. These data typically lead students to question their initial ideas and motivates the need to develop an alternative theory. This misfit between their initial expectation and data they collect also helps establish the more general enterprise as one of model fitting.

One alternative theory we prepare students to consider is based on an analysis of the set of possible outcomes — the sample space. In the Family Problem, students arrange all the 16 possible gender orders in a distribution (see Figure 17 above T_1). This distribution serves both as an *expectation* for the shape of the distributions they will later observe and also as an *explanation* for that shape. Continuing in the spirit of model-fitting, students put this new possible explanation to the test by building a computer model of the situation and conducting multiple trials to see how the simulated data are distributed (depicted below D_2 in Figure 17). In *TinkerPlots,* students can color the cases in a graph according to values on another attribute. In the graph above T_2 in Figure 17, the cases have been colored according to the specific individual outcomes. This feature helps students see the empirical distribution of the number of boys in terms of the expected distribution based on the sample space, because they can now observe that the highest frequency family of 2 boys and 2 girls is composed of 6 different specific orders while the families with 0 and 4 boys each have only one order.

In the final stages of an investigation, students focus on differences (i.e., the fit) between the distribution they expect based on the sample space and the distributions they actually get. They come to see these differences as "noise" (depicted above T_3 of Figure 17), which they can increase or decrease by changing the size of the sample they collect.

By the end of each investigation, students have observed both real data from the context and multiple repetitions of simulated data. From the multiple repetitions, the students get a better sense of sample-to-sample variability. Without the observation that samples can vary, it would not be possible in these contexts to foster a sense of data as signal and noise. Having observed the distribution of results looking different each time we press Run, we can raise the question of what, if anything, is staying the same from sample to sample. We cannot speak or think about data having a "signal" without simultaneously thinking or speaking about the noise, and vice versa. The two ideas are co-constructed.

When we consider various statistical intuitions and ideas such as the idea of signal-noise, or sample space, or the Law of Large Numbers, we tend to think of them as conceptions. But they are perceptions as well. Indeed, it is somewhat misleading to regard conception and perception as two separate, cognitive processes, as they are highly interrelated. What we know highly influences what we can see; what we see drives what we can come to know (cf. Neisser 1976).

So even though students are basically observing the same sort of data as they progress through an investigation, their conceptions and perceptions of those data are ideally changing. Once they have an image of the sample space, students are able to *see* new data they subsequently collect in a fundamentally different way. Similarly, coming to regard data as a combination of signal and noise involves perceiving data in a different way. The noise, or variability, is perhaps easiest to observe in real time, as we are collecting data. Using the computer, students can observe a distribution of sums of two dice start "growing," and, as it gets large, more closely resemble the distribution predicted from the sample space. This is not a phenomenon that we can see without preparation. Thus, another way of understanding our sequence of activities in each of the three investigations is that at each stage, we are preparing students to see something new in data collected during a subsequent stage, where the ability to "see something new" is synonymous with learning a new conception.

Without the computer, these sorts of observations are simply not possible. There is not enough classroom time to collect the amount of data needed, nor is there generally the ability to observe those data accumulating fast enough in real time so we can directly observe phenomena such as

"settling down." The Sampler comes with a control that allows students to vary the speed of the sampling from very slow to very fast. While building a model of a situation in the Sampler, students typically begin sampling at the slower speeds. This allows them to evaluate and alter the settings (such as draw and repetition numbers) until they arrive at the desired model. Once they believe their model to be correct, they on their own speed up the sampling process. At the higher speeds, they can watch the graph of its output build quickly. Most do not go to the fastest setting, which draws all the data first and only then graphs them. Rather, they choose a speed which still shows cases being added to the graph in pieces. Interestingly, these are the speeds most effective for observing the phenomenon of "settling down."

5. CONCLUDING REFLECTIONS

A major premise of our approach is that students' investigation of chance should be centered on explaining and predicting actual data. In this way, the ideas they develop about chance are not disconnected from observable phenomena. Ideally, they do not feel that they are learning a way of thinking that is divorced from their actual experiences. Having developed a possible explanation for data observed during the early part of an investigation, they then collect more data, during which they can repeatedly observe events amassing in real time.

In this report, we have offered little analysis of student thinking and learning. We have just begun an evaluation of the data collected from our last field test in which we attempted a systematic and rigorous assessment. Based on our initial impressions of the data, they include both hopeful indicators as well as some troubling ones. Among the hopeful ones are that few students at the end of the intervention seem to be reasoning according to the outcome approach (Konold, 1989). Among the troubling ones, and quite surprising to us, is how difficult it appears to be for students to learn the need to consider the sample space in thinking about new probabilistic situations. Once we remind them, most can quickly enumerate a sample space and use it to form reasonable expectations about the probability of events and the shape of distributions they are likely to observe. But without cueing, their thinking is dominated by intuitive judgments, even for situations that are isomorphic to ones we have explored together in depth. This is one reason that we have maintained three problem contexts in our materials, all of which focus on the same basic ideas. Even these three investigations, which together require

about two to three weeks of instruction, appear insufficient to prepare students to make correct predictions about similar situations without some support. However, that the amount of support required at the end of instruction is considerably less than at the beginning is an indicator of learning.

We offer this as a possible consideration: suppose most young students cannot learn the basic ideas of probabilistic thinking over a period of a few weeks. This consideration is partly based on our experiences during the past twenty years teaching data and chance and also partly on reflecting on our own learning histories. Regarding the latter, neither of us can recall a single class experience which resulted in a new insight about probability. Our development of understanding seems to be spread over many years and perhaps never gelled into a reasonably coherent system until we began teaching it ourselves. If it turned out to be the case that probabilistic thinking is only slowly acquired, this would not mean that we should put off teaching it to young students. Indeed, if probabilistic thinking is an important educational goal, and it develops slowly, then we should begin instruction as early as possible and return to the topic regularly throughout the middle and high school years. However, if core probabilistic understandings develop slowly over years rather than weeks, it poses a challenge for those of us trying to assess the impact of our instructional interventions. Time to "relearn" a concept, or the level of support needed to apply it in a new situation, may prove more sensitive indicators than items intended to demonstrate unassisted application of that concept.

6. REFERENCES

Abrahamson, D. (2006). The shape of things to come: The computational pictograph as a bridge from combinatorial space to outcome distribution. *International Journal of Computers for Mathematical Learning, 11*, 137-146.

Aspinwall, L., & Tarr, J. E. (2001). Middle school students' understanding of the role sample size plays in experimental probability. *Journal of Mathematical Behavior, 20*, 229-245.

Bakker, A., & Gravemeijer, K. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 147-168). Dordrecht, The Netherlands: Kluwer.

Ben-Zvi, D., & Arcavi, A. (2001). Junior high school students' construction of global views of data and data representations. *Educational Studies in Mathematics, 45*, 35-65.

Biehler, R. (1989). Educational perspectives on exploratory data analysis. In R. Morris (Ed.), *Studies in Mathematics Education: The Teaching of Statistics,* Vol. 7, 185-201. UNESCO: France.

Biehler, R. (1994). Probabilistic thinking, statistical reasoning, and the search for causes—Do we need a probabilistic revolution after we have taught data analysis? In J. Garfield (Ed.), *Research papers from ICOTS 4* (pp. 20-37) Minneapolis: University of Minnesota.

Biehler, R. (1997). Students' difficulties in practicing computer-supported data analysis: Some hypothetical generalizations from results of two exploratory studies. In J. B. Garfield & G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics: Proceedings of the 1996 IASE Round Table Conference* (pp. 169-190). Voorburg, The Netherlands: International Statistical Institute.

Cobb, P. (1999). Individual and collective mathematical development: The case of statistical data analysis. *Mathematical Thinking and Learning*, *1*, 5-44.

Daston, L. (1988). *Classical probability in the enlightenment*. Princeton, New Jersey: Princeton University Press.

Fischbein, E. (1975). *The intuitive sources of probabilistic thinking in children*. Dordrecht, The Netherlands: Reidel.

Hacking, I. (1975). *The emergence of probability.* London: Cambridge University Press.

Hancock, C., Kaput, J. J., & Goldsmith, L. T. (1992). Authentic inquiry with data: Critical barriers to classroom implementation. *Educational Psychologist*, *27*(3), 337-364.

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

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430-454.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction*, *6*, 59-98.

Konold, C. (1995). Confessions of a coin flipper and would-be instructor. *The American Statistician, 49*, 203-209.

Konold, C., Harradine, A., & Kazak, S. (2007). Understanding distributions by modeling them. I*nternational Journal of Computers for Mathematical Learning*, *12,* 217-230.

Konold, C., Higgins, T., Russell, S. J., & Khalil, K. (2004). *Data seen through different lenses*. Unpublished manuscript. Amherst, MA: University of Massachusetts. http://www.umass.edu/srri/serg/papers/index.html.

Konold, C., Kazak, S., Lehrer, R., & Kim, M-J. (2007). To understand a distribution, try building it from scratch. To appear in D. Pratt & J. Ainley (Eds.), *Reasoning about Informal Inferential Statistical Reasoning: A collection of current research studies. Proceedings of the Fifth International Research Forum on Statistical Reasoning, Thinking, and Literacy* (SRTL-5), University of Warwick: Warwick, UK.

Konold, C., & Lehrer, R. (2008). Technology and mathematics education: An essay in honor of Jim Kaput. In L. English (Ed.), *Handbook of International Research in Mathematics Education*, (2^{nd} edition) (pp. 49 – 72). New York: Routledge.

Konold, C., & Miller, C. (2004). *TinkerPlots™ Dynamic Data Exploration* (Version 1.0). Emeryville, CA: Key Curriculum Press.

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

Konold, C., Pollatsek, A., Well, A., & Gagnon, A. (1997). Students analyzing data: Research of critical barriers. In J. B. Garfield & G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics: Proceedings of the 1996 IASE Round Table Conference* (pp. 151-167). Voorburg, The Netherlands: International Statistical Institute.

Landwehr, J. M. (1985). Using microcomputers for data analysis and simulation experiments in junior and senior high school. In L. Råde & T. Speed, (Eds.). *Teaching of statistics in the computer age. Proceedings of the Sixth ISI Round Table Conference on the Teaching of Statistics* (pp. 105-113). Bromley, Kent: Chartwell Bratt Ltd.

Lecoutre, M. P. (1992). Cognitive models and problem spaces in "purely random" situations. *Educational Studies in Mathematics, 23*, 557-568.

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 in Mathematics Education, 12*, 195-216.

Lehrer, R., Konold, C., & Kim, M-J. (2006). *Constructing data, modeling chance in the middle school*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Lehrer, R., & Schauble, L. (2000). Modeling in mathematics and science. In R. Glaser (Ed.), Advances in Instructional Psychology: Vol. 5. Educational Design and Cognitive Science (pp. 101-159). Mahwah, NJ: Erlbaum.

Makar, K., & Rubin, A. (2007). Beyond the bar graph: Teaching informal statistical inference in primary school. In Ainley, J. and Pratt, D. (Eds.) *Reasoning about Statistical Inference: Innovative Ways of Connecting Chance and Data*, The University of Warwick, 11-17 November 2007.

Masnick, A. M., Klahr, D. & Morris, B. J. (2007). Separating signal from noise: Children's understanding of error and variability in experimental outcomes. In M. C. Lovett & P. Shah (Eds.), *Thinking with Data* (pp. 3-26). New York: Taylor & Francis.

Moore, D. S. (1992). Teaching statistics as a respectable subject. In F. S. Gordon & S. P. Gordon (Eds.), *Statistics for the twenty- first century*, (MAA Notes, #26, pp. 14-25). Mathematical Association of America.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematic*s. Reston, VA: National Council of Teachers of Mathematics.

Neisser, U. (1976). *Cognition and reality*. San Francisco: Freeman.

Newman, C. M., Obremski, T.E., & Scheaffer, R.L. (1987). *Exploring Probability*, Quantitative Literacy Series. Palo Alto, CA: Dale Seymour.

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

Pfannkuch, M. (2006). Informal inferential reasoning. In A. Rossman & B. Chance (Eds.) *Proceedings of the 7th International Conference on Teaching Statistics (ICOTS)* [CD-ROM]. Salvador, Bahai, Brazil.

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: Kluwer Academic Publishers.

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children*. New York: W. W. Norton & Company Inc.

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

Pratt, D. (2000). Making sense of the total of two dice. *Journal of Research in Mathematics Education, 31*, 602-625.

Rubin, A., Hammerman, J., & Konold, C. (2006). Exploring informal inference with interactive visualization software. In A. Rossman & B. Chance (Eds.) *Proceedings of the 7th International Conference on Teaching Statistics (ICOTS)* [CD-ROM]. Salvador, Bahai, Brazil.

Scheaffer, R. L., Watkins, A. E., & Landwehr, J. M. (1998). What every high-school graduate should know about statistics. In S. P. LaJoie (Ed.), *Reflections on statistics: Agendas for learning, teaching, and assessment in K-12* (pp.3-318). Mahwah, NJ: Lawrence Erlbaum.

Sedlmeier, P. (2007). Statistical reasoning: Valid intuitions put to use. In M. C. Lovett & P. Shah (Eds.), *Thinking with Data* (pp. 389-419). New York: Taylor & Francis.

Sedlmeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making, 10*, 33-51.

Steinbring, H. (1991). The theoretical nature of probability in the classroom. In R. Kapadia & M. Borovcnik (Eds.), *Chance encounters: Probability and education* (pp. 135-176). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Shaughnessy, J.M. (1997). Missed opportunities in research on the teaching and learning of data and chance. In F. Biddulph and K. Carr (Eds.), *People in mathematics education* (Vol. 1, pp. 6-22). Waikato, New Zealand: Mathematics Education Research Group of Australasia.

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

Stohl, H., & Tarr, J. E. (2002). Developing notions of inference with probability simulation tools. *Journal of Mathematical Behavior, 21(3)*, 319-337.

Tarr, J. E., Lee, H. S., & Rider, R. L. (2006). When data and chance collide: Drawing inferences from empirical data. In G. F. Burrill & P. C. Elliott (Eds.), *Thinking and reasoning with data and chance: Sixty-eight Yearbook* (pp. 139-149). Reston, VA: The National Council of Teachers of Mathematics.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Watkins, A., Burrill, G., Landwehr, J., & Schaeffer, R. (1992). Remedial statistics?: The implications for colleges of the changing secondary school curriculum. In F. S. Gordon & S. P. Gordon (Eds.), *Statistics for the twenty- first century*, (MAA Notes, #26, pp. 45-55). Mathematical Association of America.

Watson, J.M., Collis, K. F., Callingham, R. A., & Moritz, J. B. (1995). A model for assessing higher order thinking in statistics. *Educational Research and Evaluation*, *1(3),* 247-275.

Watson, J. M., & Moritz, J. B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics*, *37*, 145-168.

Watson, J. M., & Moritz, J. B. (2003). Fairness of dice: A longitudinal study of students' beliefs and strategies for making judgments. *Journal for Research in Mathematics Education, 34 (4),* 270-303.

Well, A. D., Pollatsek, A., & Boyce, S. J. (1990). Understanding the effects of sample size on the variability of the mean. *Journal of Organizational Behavior and Human Decision Processes, 47*, 289-312.

Wild, C. (2006). The concept of distribution. *Statistics Education Research Journal*, *5(2)* 10-26.

Wilensky, U. (1997). What is normal anyway? Therapy for epistemological anxiety. *Educational Studies in Mathematics, 33*, 171-202.