Data Analysis as the Search for Signals in Noisy Processes

Clifford Konold and Alexander Pollatsek, *University of Massachusetts, Amherst*

The idea of data as a mixture of signal and noise is perhaps the most fundamental concept in statistics. Research suggests, however, that current instruction is not helping students to develop this idea, and that though many students know, for example, how to compute means or medians, they do not know how to apply or interpret them. Part of the problem may be that the interpretations we often use to introduce data summaries, including viewing averages as typical scores or fair shares, provide a poor conceptual basis for using them to represent the entire group for purposes such as comparing one group to another. To explore the challenges of learning to think about data as signal and noise, we examine the "signal/noise" metaphor in the context of three different statistical processes: repeated measures, measuring individuals, and dichotomous events. On the basis of this analysis, we make several recommendations about research and instruction.

Key Words: Conceptual knowledge; Curriculum; Historical analysis; Reform in mathematics education; Statistics

Until recently, the study of statistics in the United States was confined to the university years. Following recommendations made by the National Council of Teachers of Mathematics (NCTM, 1989; 2000), and building on the ground-breaking Quantitative Literacy series (see Scheaffer, 1991), statistics and data analysis are now featured prominently in most mathematics curricula and are also appearing in the K–12 science standards and curricula (Feldman, Konold, & Coulter, 2000; National Research Council, 1996). Concurrently, university-level introductory statistics courses are changing (e.g., Cobb, 1993; Gordon & Gordon, 1992; Smith, 1998) in ways that pry them loose from the formulaic approach copied with little variation in most statistics textbooks published since the 1950s.[1] At all levels, there is a new commitment to involve students in the analysis of real data to answer practical questions. Formal inference, at the introductory levels, is taking a less prominent place as greater emphasis is given to exploratory approaches (à la Tukey, 1977) to reveal structure in data. This approach often capitalizes on

---

[1] As George Cobb (1993) remarked, "If one could superimpose maps of the routes taken by all elementary books, the resulting picture would look much like a time-lapse night photograph of car taillights all moving along the same busy highway" (¶ 53).

---

the power of visual displays and new graphic-intensive computer software (Biehler, 1989; Cleveland, 1993; Konold, 2002).

Despite all the criticisms that we could offer of the traditional introductory statistics course, it at least has a clear objective: to teach ideas central to statistical inference, including the Law of Large Numbers and the Central Limit Theorem. For the students now learning more exploratory forms of data analysis, the objective is less clear. There are various proposals about which core ideas we should target in early instruction in data analysis. Wild and Pfannkuch (1999), for example, view variation as the core idea of statistical reasoning and propose various subconstructs that are critical to learning to reason about data. Recently designed and tested materials for 12- to 14-year-olds aim at developing the idea of a distribution (Cobb, 1999; Cobb, McClain, & Gravemeijer, in press). According to the supporting research, this idea entails viewing data as "entities that are distributed within a space of possible values," in which various statistical representations—be they types of graphical displays or numerical summaries—are viewed as different ways of structuring or describing distributions (see Cobb, 1999, pp. 10–11). Others have argued the centrality of the idea of data as an aggregate—an emergent entity (i.e., distribution) that has characteristics not visible in any of the individual elements in the aggregate (Konold & Higgins, in press; Mokros & Russell, 1995).

In this article, we build on these ideas of variation, distribution, and aggregate to offer our own proposal for the core idea that we believe should guide statistics and data analysis instruction, beginning perhaps as early as age 8. In short, that idea involves coming to see statistics as the study of noisy processes—processes that have a signature, or signal, which we can detect if we look at sufficient output.

It might seem obvious that a major purpose of computing statistics such as the mean or median is to represent such a "signal" in the "noise" of individual data points. However, this idea is virtually absent from our curricula and standards documents. Neither NCTM's *Principles and Standards for School Mathematics* (2000) nor AAAS's *Science for All Americans* (1989) explicitly describes an average as anything like a signal. Our search through several middle school and high school mathematics curricula has not uncovered a single reference to this idea. Nor does it appear in earlier research investigating students' ideas about averages and their properties (Mokros & Russell, 1995; Pollatsek, Lima, & Well, 1981; Strauss & Bichler, 1988). The idea is evident, however, in a few recent studies. In their investigation of statistical reasoning among practicing nurses, Noss, Pozzi, and Hoyles (1999) refer briefly to this interpretation and quote one nurse whom they interviewed as characterizing a person's average blood pressure as "what the normal range was sort of settling down to be." The idea of signal and noise is also evident in the work of Biehler (1994), Wild and Pfannkuch (1999), and Wilensky (1997).

OVERVIEW

We begin by describing how statisticians tend to use and think about averages as central tendencies. We then contrast this interpretation with various other inter-

pretations of averages that we frequently encounter in curriculum materials. Too frequently, curricula portray averages as little more than summaries of groups of values.[2] Although this approach offers students some rationale for summarizing group data (for example, to see what is "typical"), we will argue that it provides little conceptual basis for using such statistical indices to characterize a set of data, that is, to represent the whole set. To support this claim, we review research that has demonstrated that although most students know how to compute various averages such as medians and means, few use averages to represent groups when those averages would be particularly helpful—to make a comparison between two groups. We recommend beginning early in instruction to help students develop the idea of central tendency (or data as a combination of signal and noise). To explore the conceptual underpinnings of the notion of central tendency, we briefly review its historical development and then examine three types of statistical processes. For each process, we evaluate the conceptual difficulty of regarding data from that process as a combination of signal and noise. Finally, we outline some possible directions for research on student thinking and learning.

In this article, we focus our discussion on averages, with an emphasis on means (using the term "average" to refer to measures of center collectively, including the mean, median, and mode). By focusing on averages, we risk being misunderstood by those who have recently argued that instruction and public discourse have been overemphasizing measures of center at the expense of variability (e.g., Shaughnessy, Watson, Moritz, & Reading, 1999; also, see Gould, 1996). A somewhat related but more general critique comes from proponents of Tukey's (1977) Exploratory Data Analysis (EDA), who advocate that, rather than structure our curricula around a traditional view of inferential statistics, we should instruct young students in more fluid and less theory-laden views of analysis (e.g., Biehler, 1989; 1994).

Those concerned that measures of center have been overemphasized as well as proponents of EDA may misread us as suggesting that instruction should aim at teaching students to draw conclusions by inspecting a limited number of simple summaries such as means. In fact, we agree wholeheartedly with Shaughnessy et al. (1999) and with EDA proponents that we should be teaching students to attend to general distributional features such as shape and spread, and to look at distributions in numerous ways for insights about the data. We do not view the decision to focus our analysis here on measures of center as being at odds with their concerns. Our decision is partly pragmatism and partly principle.

---

[2] David Krantz (personal communication, December 13, 2001) shared with us his response to the question "Do we really need the mean in descriptive stats?" which had appeared on a data analysis listserv. "I'm not very clear on what is meant by 'descriptive statistics.' To be honest, I don't think there is any such thing, except as a textbook heading to refer to the things that are introduced prior to consideration of sampling distributions. Any description must have a purpose if it is to be useful—it is supposed to convey something real. The line between 'mere description' and suggesting some sort of inference is very fuzzy."

On the pragmatic side, we wanted to simplify our exposition. Almost all statistical measures capture group properties, and they share an important property with good measures of centers: they stabilize as we collect more data. Those measures include those of spread, such as the standard deviation, interquartile range, percentiles, and measures of skewness. But switching among these different measures would needlessly complicate our exposition.

The deeper reason for focusing our discussion on measures of center is that we believe such measures do have a special status, particularly for comparing two sets of data, and here some proponents of teaching EDA may well disagree with us. Biehler (1994), for example, maintained that the distribution should remain the primary focus of analysis and that we should regard an average, such as the mean, as just one of many of its properties. We will argue that the central idea should be that of searching for a signal, and that the idea of distribution comes into better focus when it is viewed as the "distribution around" a signal. Furthermore, we claim that the most basic questions in analyzing data involve looking at *group differences* to determine whether some factor has produced a difference in the two groups. Typically, the most straightforward and compelling way to answer these questions is to compare averages. We believe that much of statistical reasoning will elude students until they understand when a comparison of two averages makes sense and, as a corollary, when such a comparison is misleading. If they do not understand this, students' explorations of data (i.e., "data snooping") will almost certainly lack direction and meaning.

SIGNALS IN NOISY PROCESSES

A statistician sees group features such as the mean and median as indicators of stable properties of a variable system—properties that become evident only in the aggregate. This stability can be thought of as the certainty in situations involving uncertainty, the signal in noisy processes, or, the descriptor we prefer, central tendency. Claiming that modern-day statisticians seldom use the term "central tendency," Moore (1990, p. 107) suggests that we abandon the phrase and speak instead of measures of "center" or "location." But we use the phrase here to emphasize conceptual aspects of averages that we fear are often lost, especially to students, when we talk about averages as if they were simply locations in distributions.

By central tendency we refer to a *stable* value that (a) represents the signal in a variable process and (b) is better approximated as the number of observations grows.[3] The obvious examples of statistics used as indicators of central tendency are averages such as the mean and median. Processes with central tendencies have two components: (a) a stable component, which is summarized by the mean, for example, and (b) a variable component, such as the deviations of individual scores around an average, which is often summarized by the standard deviation.

---

[3] Many use the term *central tendency* as a synonym for *average* or *center*. When we refer to central tendency in this article, we have in mind the particular definition specified here.

It is important to emphasize that measures of center are not the only way to characterize stable components of noisy processes. The shape of a frequency distribution as well as global measures of variability, for example, also stabilize as we collect more data; they, too, give us information about the process. We might refer to this more general class of characteristics as *signatures* of a process. We should point out, however, that all the characteristics that we might look at, including the shape and variability of a distribution, are close kin to averages. That is, when we look at the shape of a particular distribution, we do not ordinarily want to know precisely how the frequency of values changes over the range of the variable. Rather, we tame the distribution's "bumpiness." We might do this informally by visualizing a smoother underlying curve or formally by computing a best-fit curve. In either case, we attempt to see what remains when we smooth out the variability. In a similar manner, when we employ measures such as the standard deviation or interquartile range, we strive to characterize the *average* spread of the data in the sample.

Implicit in our description of central tendency is the idea that even as one speaks of some stable component, one acknowledges the fundamental variability inherent in that process and thus its probabilistic nature. Because of this, we claim that the notion of an average understood as a central tendency is inseparable from the notion of spread. That average and variability are inseparable concepts is clear from the fact that most people would consider talking about the average of a set of identical values to be odd. In addition, it is hard to think about why a particular measure of center makes sense without thinking about its relation to the values in the distribution (e.g., the mean as the *balance point* around which the sum of the deviation scores is zero, or the median as the point where the number of values above equals the number of values below).

Not all averages are central tendencies as we have defined them above. We could compute the mean weight of an adult lion, a Mazda car, and a peanut, but no clear process would be measured here that we could regard as having a central tendency. One might think that the mean weight of all the lions in a particular zoo would be a central tendency. But without knowing more about how the lions got there or their ages, it is questionable whether this mean would necessarily tell us anything about a process with a central tendency. Quetelet described this distinction in terms of *true* means of distributions that follow the law of errors versus *arithmetic* means that can be calculated for any assortment of values, such as our hodgepodge above (see Porter, 1986, p. 107).

Populations Versus Processes

In the preceding description, we spoke of processes rather than populations. We contrast these two ways of thinking about samples or batches of data, as shown in Figure 1. When we think of a sample as a subset of a population (see the left graphic), we see the sample as a piece allowing us to guess at the whole: the average and shape of the sample allow us perhaps to estimate the average and shape of the population. If we wanted to estimate the percentage of the U.S. population favoring

A sample from a population   Output from a noisy process
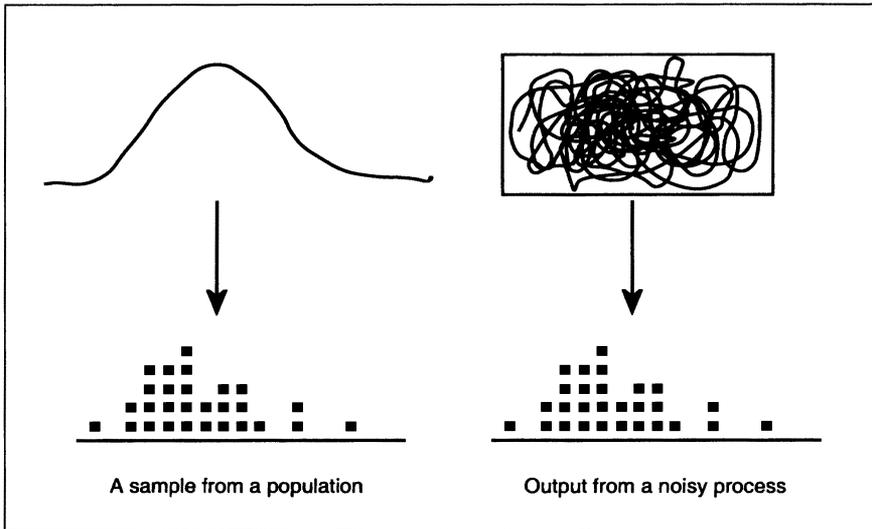
Figure 1. Data viewed as a sample of a population (left) versus data viewed as output of a noisy process (right).

gun control, we would imagine there being a population percentage of some unknown value, and our goal would be to estimate that percentage from a well-chosen sample. Thinking in these terms, we tend to view the population as static and to push to the background questions about why the population might be the way it is or how it might be changing.

From the process perspective (as depicted in the right graphic of Figure 1), we think of a population or a sample as resulting from an ongoing, dynamic process, a process in which the value of each observation is determined by a large number of causes, some of which we may know and others of which we may not. This view moves to the foreground questions about why a process operates as it does and what factors may affect it. In our gun control example, we might imagine people's opinions on the issue as being in a state of flux, subject to numerous and complex influences. We sample from that process to gauge the net effect of those influences at a point in time, or perhaps to determine whether that process may have changed over some time period.

For many of the reasons discussed by Frick (1998), we have come to prefer thinking of samples (and populations, when they exist) as outputs of processes.[4]

---

[4] Adopting this perspective, we will generally refer to *processes* rather than to *populations*, to *signals* or *central tendencies of processes* rather than to *population parameters*, and to *estimates of signals* rather than to *sample statistics*. We use the term *process* to refer both to processes that remain relatively stable over time as well as to stochastic processes, which can change quickly over time.

One reason for this preference is that a process view better covers the range of statistical situations in which we are interested, many of which have no real population (e.g., weighing an object repeatedly). Another reason for preferring the process view is that when we begin thinking, for example, about how to draw samples, or why two samples might differ, we typically focus on factors that play a role in producing the data. That is, we think about the *causal processes* underlying the phenomena we are studying. Biehler (1994) offered a similar analysis of the advantages of viewing data as being produced by a probabilistic mechanism—a mechanism that could be altered to produce predictable changes in the resultant distribution. Finally, viewing data as output from a process highlights the reason that we are willing to view a collection of individual values as in some sense "the same" and thus to reason about them as a unity: We consider them as having been generated by the same process.

This notion of process is, of course, inherent in the statistician's conception of a population, and we expect that most experts move between the process and population perspectives with little difficulty or awareness.[5] However, for students new to the study of statistics, the choice of perspective could be critical. To illustrate more fully what we mean by reasoning about processes and their central tendencies, we discuss recent results of the National Assessment of Educational Progress (NAEP).

NAEP Results as Signals of Noisy Processes

NAEP is an assessment of student capabilities in Grades 4, 8, and 12, conducted every four years in the United States. On the 1998 assessment, eighth graders averaged 264 on the reading component.[6] What most people want to know, of course, is how this compares to the results from previous assessments. In this case, the mean had increased 4 points since the 1994 assessment. The twelfth graders had also gained 4 points on average since 1994, and the fourth graders, 3 points. Donahue, Voelkl, Campbell, and Mazzeo (1999) interpreted these differences as evidence that children's reading scores were improving.

Reports such as this are now so commonplace that we seldom question the logic of this reasoning. But what is the rationale in this case for comparing group means and for taking the apparently small difference between those means seriously? We will argue that to answer these questions from a statistical perspective requires a well-formed idea of a central tendency.

Interpreted as a central tendency, the mean of 264 is a measure of a complex process that determines how well U.S. children read at a given point in time. An obvious component of this process is the reading instruction that children receive

---

[5] However, Frick (1998) argues that the difference between processes and populations is more than terminology, claiming that the tension between theoretical descriptions of random sampling and what we actually do in practice could be resolved if we thought explicitly of sampling from processes rather than from populations.

[6] The maximum score on the reading component was 500, and the standard deviation was 50.

in school. Another component of the process is the behavior of adults in the home, such as their personal reading habits, the time that they spend reading to their children, and the kind and quantity of reading material that they have in the home. A third component consists of factors operating outside the home and school, including determinants of public health and development, such as nutrition levels and the availability and use of prenatal care, genetic factors, and the value placed on literacy and education by local communities and the society at large.

Using a statistical perspective, we often find it useful to regard all these influences together (along with many others that we may be unaware of) as a global process that turns out readers of different capabilities. In the sense that we cannot know how these various factors work together in practice to produce results, the global process is a probabilistic one, unpredictable at the micro level. However, even though readers produced by this process vary unpredictably in their performance, we can regard the entire process at any given point in time as having a certain stable capability to produce competent readers. The average performance of a large sample of readers produced by this process is one way to gauge the power of that process (or its propensity) to produce a literate citizenry. As Mme de Staël explained in 1820, "events which depend on a multitude of diverse combinations have a periodic recurrence, a fixed proportion, when the observations result from a large number of chances" (as quoted in Hacking, 1990, p. 41). And because of the convergence property of central tendencies, the larger the data set, the better the estimate we expect our sample average to be of the stable component of the process.

Given the huge sample size in the reading example (about 11,000 eighth graders) and assuming proper care in composing the sample, we expect that the sample mean of 264 is very close to this propensity. Assuming that the 1994 mean is of equal quality, we can be fairly certain that the difference between these two means reflects a real change in the underlying process that affects reading scores. Note that the important inference here does not concern a sampling issue in the narrow sense of randomly sampling from a fixed known population. That is, assuming no changes in the system, we would expect next year's mean to come out virtually the same even though the population of eighth graders would consist of different individuals. Focusing on the process rather than the population helps make the real intent of our question clear.

The mean is not necessarily the best single number to serve as an index of such a change. The median is also a good index, and changes in the 25th percentile, the percent above some minimal value, the standard deviation, or the interquartile range could also be valid indicators of changes in the underlying educational process. As long as a process remains stable, we expect the mean, or any of these other statistical indices obtained from that process, to remain relatively unchanged from sample to sample. Conversely, when a statistic from a large sample changes appreciably, we assume that the process has changed in some way. Furthermore, these expectations are crucial in our attempts to evaluate efforts to alter processes. In the case of reading, we might introduce new curricula, run an advertising campaign encouraging parents to read to their children, expand the school free lunch program

in disadvantaged areas, and upgrade local libraries. If we do one or more of these things and the mean reading scores of an appropriate sample of children increases, we have grounds for concluding that we have improved the process for producing readers. Again, we emphasize that though we have specified the mean in this example, we might be as happy using the median or some other measure of center.

The above example, however, indicates a way in which a measure of center is often special. That is, the practical issue in which we are usually interested is whether, overall, things are getting better or worse, a question most naturally phrased in terms of a change of center. It is much harder to think of examples where we *merely* want to increase or decrease the variability or change the shape of the distribution. We could imagine an intervention that tried only to narrow the gap between good and poor readers, in which case we would compare measures of spread, such as the standard deviation. Although there are questions that are naturally phrased in terms of changes in variability or distribution shape, such questions are typically second-order concerns. That is, we usually look at whether variability or shape have changed to determine whether we need to qualify our conclusion about comparing measures of center. Even in situations where we might be interested in reducing variability, such as in income, we are certainly also interested in whether this comes at the expense of lowering the average.

DIFFERENT INTERPRETATIONS OF AVERAGES

We have argued that statisticians view averages as central tendencies, or signals in variable data. But this is not the only way to think about them. In Table 1, we list this interpretation along with several others, including viewing averages as data reducers, fair shares, and typical values. We consider an interpretation to be the goal that a person has in mind when he or she computes or uses an average. It is the answer that a person might give to the question "Why did you compute the average of those values?" Some of these interpretations are described in Strauss and Bichler (1988) as "properties" of the mean. Mokros and Russell (1995) described other interpretations as "approaches" that they observed elementary and middle school students using.[7] In Table 1, we also provide an illustrative problem context for each interpretation. Of course, any problem could be interpreted from a variety of perspectives. But we chose these particular examples because their wording seemed to suggest a particular interpretation.

Data Reduction

According to this view, averaging is a way to boil down a set of numbers into one value. The data need to be reduced because of their complexity—in particular, because of the difficulty of holding the individual values in memory. Freund and

---

[7] See Bakker (2001) for a review of the historical origins of various types of averages and a discussion of parallels between these ideas and the development of student thinking.

Table 1
Examples of Contexts for Various Interpretations of Average

Interpretation/ meaning	Example context
Data reduction	Ruth brought 5 pieces of candy, Yael brought 10 pieces, Nadav brought 20, and Ami brought 25. Can you tell me in one number how many pieces of candy each child brought? (From Strauss & Bichler, 1988)
Fair share	Ruth brought 5 pieces of candy, Yael brought 10 pieces, Nadav brought 20, and Ami brought 25. The children who brought many gave some to those who brought few until everyone had the same number of candies. How many candies did each girl end up with? (Adapted from Strauss & Bichler, 1988)
Typical value	The numbers of comments made by 8 students during a class period were 0, 5, 2, 22, 3, 2, 1, and 2. What was the typical number of comments made that day? (Adapted from Konold & Garfield, 1992.)
Signal in noise	A small object was weighed on the same scale separately by nine students in a science class. The weights (in grams) recorded by each student were 6.2, 6.0, 6.0, 15.3, 6.1, 6.3, 6.2, 6.15, 6.2. What would you give as the best estimate of the actual weight of this object? (Adapted from Konold & Garfield, 1992.)

Wilson (1997) draw on this interpretation to introduce averages in their text: "Although distributions provide useful descriptions of data, they still contain too much detail for some purposes" (p. 15). They characterize numerical summaries as ways to further simplify data, warning that "this condensation or data reduction may be accompanied by a loss of information, such as information on the shape of the distribution" (p. 16). One of the high school students interviewed by Konold, Pollatsek, Well, and Gagnon (1997) used this as a rationale for why she would look at a mean or median to describe the number of hours worked by students at her school.

> We could look at the mean of the hours they worked, or the median…. It would go through a lot to see what every, each person works. I mean, that's kind of a lot, but you could look at the mean…. You could just go through every one … [but] you're not going to remember all that.

Fair Share

The computation for the mean is often first encountered in elementary school in the context of fair-share problems, with no reference to the result being a mean or average. Quantities distributed unevenly among several individuals are collected and then redistributed evenly among the individuals. The word *average*, in fact, derives from the Arabic *awariyah*, which translates as "goods damaged in shipping." According to Schwartzman (1994), the Italians and French appropriated this term to refer to the financial loss resulting from damaged goods. Later, it came to specify the portion of the loss borne by each of the many people who invested in

the ship. Strauss and Bichler (1988) provided eleven problems as examples of tasks that they used in their research, and we would regard all but three of them as involving the idea of fair share. We can view many commonly encountered rates, such as yearly educational expenditure per student, as based on the fair-share idea, since we tend to think most naturally about these rates as distributing some total quantity equally over some number of units. In such cases, we do not ordinarily think of the computed value in relation to each individual value, nor do we worry, when computing or interpreting this fair share, about how the component values are distributed or whether there are outliers.

Typical Value

Average as a typical score is one of the more frequently encountered interpretations in current precollege curricula. What appears to make values typical for students are their position (located centrally in a distribution of values) and/or their frequency (being the most frequent or even the majority value). Younger students favor the mode for summarizing a distribution presumably because it can often satisfy both of these criteria (Konold & Higgins, in press). Mokros and Russell (1995) speculated that those students they interviewed who used only modes to summarize data may have interpreted *typical* as literally meaning the most frequently occurring value. Researchers have also observed students using as an average a range of values in the center of a distribution (Cobb, 1999; Konold, Robinson, Khalil, Pollatsek, Well, Wing, & Mayr, 2002; Mokros & Russell, 1995; Noss, Pozzi, & Hoyles, 1999; Watson & Moritz, 1999). These "center clumps" are located in the heart of the distribution and often include a majority of the observations. In this respect, these clumps may serve as something akin to a mode for some students.

Signal in Noise

According to this perspective, each observation is an estimate of an unknown but specific value. A prototypical example is repeatedly weighing an object to determine its actual weight. Each observation is viewed as deviating from the actual weight by a measurement error, which is viewed as "random." The average of these scores is interpreted as a close approximation to the actual weight.

Formal Properties of Averages

Many school tasks involving averages seem unrelated to any of the particular interpretations we describe above. For example, finding the average of a set of numbers out of context seems intended only to develop or test students' computational abilities. Other school tasks explore formal properties of averages, which we also would not view as directly related to particular interpretations. Such tasks include those meant to demonstrate or assess the idea that (a) the mean of a set of numbers is simply related to the sum of those numbers, (b) the mean is a balance

point and the median a partition that divides the cases into two equal-sized groups,[8] (c) the mean and median lie somewhere within the range of the set of scores and (d) the mean or median need not correspond to the value of an actual observation. In their longitudinal study of the development of young students' understandings of average, Watson and Moritz (2000) focused in particular on these relations, asking students, for example, how the mean number of children per family could possibly be 2.3 rather than a whole number. We consider most of the properties enumerated by Strauss and Bichler (1988, p. 66) to be formal relations of this sort. We are not arguing that these are unimportant or trivial ideas but rather that they are usually not tied to particular interpretations of averages.

Applying Interpretations to the Problem of Group Comparison

In the NAEP example, we explored the notion of central tendency and showed how it provides a basis for using averages—means, in that case—to compare groups. Owing to the fact that the mean is a very stable estimator in large samples, we can use it to track changes in a process even though the output from that process is variable and unpredictable in the short run.

What bases do the other interpretations of average provide for evaluating the two NAEP results by comparing means? Consider the data reduction interpretation: Data are distilled to a single value, presumably because of our inability to consider all the values together. We argue that nothing in this interpretation suggests that any new information emerges from this process; indeed, a considerable loss of information seems to be the price paid for reducing complexity. By this logic, it would seem that as a data set grows larger, any single-value summary becomes less representative of the group as increasingly more information is lost in the reduction process.

The typical-value interpretation is nearer to the central tendency interpretation in that it may involve the idea that the value, in some sense, represents much of the data in the group. However, as with the data reduction interpretation, it is not clear why one ideally would like to have typical values from large samples rather than from small ones. Indeed, it would seem as reasonable to regard a typical score as becoming less (rather than more) representative of a group as that group became larger and acquired more deviant values.

The fair-share interpretation may provide some basis for using means to compare groups. One could think of the mean in the 1998 NAEP data as the reading score that all students sampled that year would have if reading ability were divided evenly among all the students sampled. Based on this reasoning, one might reasonably conclude that the 1998 group had a higher reading score than the 1994 group.

---

[8] There are good grounds for considering the idea of mean as balance point as an interpretation. This interpretation figures centrally in mechanics, where the mean is a measure of center of mass. But in the statistics texts that we examined, the idea of mean as balance point seemed to be used solely as a way to visualize the location of the mean in a distribution of values and not as an interpretation as we have defined it.

Cortina, Saldanha, and Thompson (1999) explored the use of this notion by seventh- and eighth-grade students and concluded that these students could use the idea of fair share to derive and compare means of unequal groups. However, we would guess that many students would regard such reasoning skeptically unless it were physically possible to reallocate quantities in the real-world situation. If, for example, we were thinking about the number of boxes of cookies sold by different scout troops (as in the study by Cortina et al.), redistributing the cookie boxes evenly makes some sense. In contrast, if we were reasoning about mean weight, height, or IQ of a number of individuals, we would have to think of these pounds, inches, or IQ points being shared metaphorically.[9]

Furthermore, we are skeptical about whether the fair-share interpretation is a statistical notion at all. It seems to ignore, in a sense, the original distribution of values and to attend only to the total accumulation of some amount in a group. Consider, for example, the value we would compute to decide how the different numbers of candies brought by various children to a party could be equally redistributed among the children (see Table 1). In this context, the particulars about how the candies were originally distributed seems irrelevant. That is, the number that constitutes a fair share is not viewed as a representation or summary of the original distribution but rather as the answer to the question of how to divide the candies equitably.

In conclusion, whereas some of the interpretations may be useful to summarize a group of data, it is quite another thing to take a statistic seriously enough as to use it to represent the *entire* group as one must do when using averages to compare groups. We claim that viewing an average as a central tendency provides a strong conceptual basis for, among other things, using averages to compare two groups, whereas various other interpretations of average, such as data reducers and typical values, do not.

We acknowledge that our analysis of these alternative interpretations has been cursory and that it therefore should be regarded skeptically. However, our primary purpose is to highlight some of the questions that we should ask in exploring different approaches to introducing students to averages. Furthermore, there is good evidence that whatever interpretations students do have of averages, those interpretations usually do not support using averages to compare one group to another. Many studies have demonstrated that even those who know how to compute and use averages in some situations do not tend to use them to compare groups.

Students' Tendency Not to Use Averages to Compare Groups

Gal, Rothschild, and Wagner (1990) interviewed students of age 8, 11, and 14 about their understanding of how means were computed and what they were useful for. They also gave the students nine pairs of distributions in graphic form and asked

---

[9] We have to be careful using this logic. For example, mean income would be a different, and probably better, indicator of the power of the economic system to take care of its citizens if the wealth were, in fact, distributed equally.

them to decide whether the groups were different or not. Only half of the 11- and 14-year-olds who knew how to compute the mean of a single group (and, also, to some extent, how to interpret it) went on to use means to compare two groups. Hancock, Kaput, and Goldsmith (1992) and, more recently, Watson and Moritz (1999) have reported similar findings.

This difficulty is not limited to the use of means. Bright and Friel (1998) questioned 13-year-old students about a stem-and-leaf plot that showed the heights of 28 students who did not play basketball. They then showed them a stem-and-leaf plot that included this data along with the heights of 23 basketball players. This latter plot is shown in Figure 2. Heights of basketball players were indicated in bold type, as they are here. Students had learned how to read this type of display and had no difficulty reading values from it. Asked about the "typical height" in the single distribution of the non–basketball players, the students responded by specifying middle clumps (e.g., 150–160 cm), a reasonable group summary. Yet, shown the plot with both distributions, they could not generalize this method or find another way to determine "How much taller are the basketball players than the students who did not play basketball?"

We found similar difficulties when we interviewed four high school seniors (ages 17–18) who had just completed a yearlong course in probability and statistics

```
10
11
12
13    8  8  8  9
14    1  2  4  7  7  7
15    0  0  1  1  1  1  2  2  2  2  3  3  5  6  6  7  8
16
17    1
18    0  3  5
19    0  2  5  7  8  8
20    0  0  2  3  5  5  5  5  7
21    0  0  0  5
22    0
23
```
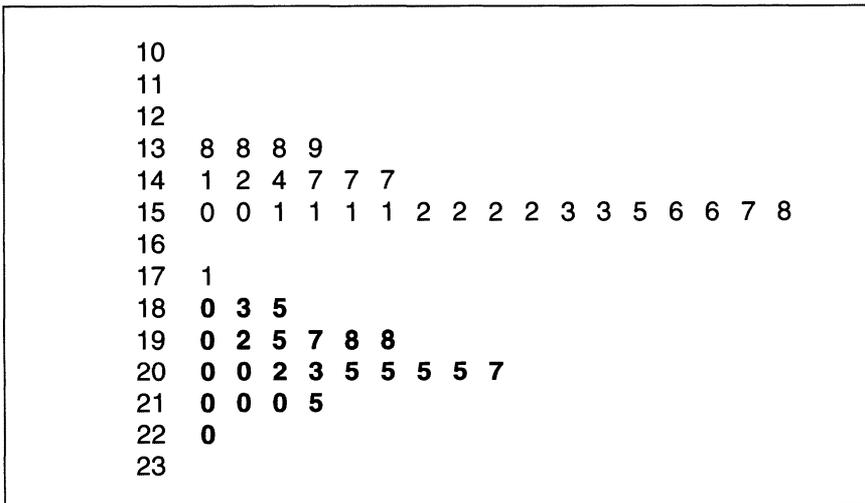
Figure 2. Stem-and-leaf plot of heights of students and basketball players (**boldface**) from "Helping Students Interpret Data," by G. Bright and S. N. Friel, in *Reflections on Statistics: Learning, Teaching, and Assessment in Grades K–12* (p. 81), edited by S. P. Lajoie, 1998, Mahwah, NJ: Lawrence Erlbaum Associates. Copyright 1998 by Lawrence Erlbaum Associates.

Note. The row headed by 13 (the stem) contains four cases (leaves)—three students of 138 cm and a fourth student of 139 cm.

(Biehler, 1997; Konold et al., 1997). During the course, they had frequently used medians (primarily in the context of box plot displays) as well as means to make group comparisons. However, during a postcourse interview in which they were free to use whatever methods of comparison seemed appropriate, they seldom used medians or means for this purpose. Instead, they tended to compare the number of cases in each group that had the same value on the dependent variable. For example, to decide if males were taller than females, they might inspect the sample for all individuals who were 6 feet tall and argue that males were taller because there were more males than females of that height. In making these comparisons, they typically did not attend to the overall number of individuals in the two groups (in this case, to the overall number of males vs. females). Other researchers, including Cobb (1999) and Watson and Moritz (1999), have reported students using this same "slicing" technique over a range of different problems to compare two groups.

In short, despite the fact that instruction in statistics usually focuses on averages, many students do not use those measures when they would be particularly helpful— to make comparisons between groups composed of variable elements. We suggest that this pattern is symptomatic of students' failure to interpret an average of a data set as saying something about the entire distribution of values. To address this problem instructionally, we believe that we should be encouraging students early in statistics instruction to think of averages as central tendencies or signals in noisy processes. We acknowledge that this is a complex idea and one that is particularly difficult to apply to the type of processes that we often have students investigating. We explore these conceptual difficulties below.

THREE TYPES OF PROCESSES
AND THEIR CONCEPTUAL CHALLENGES

Hints about the cognitive complexity of central tendency are found in the historical account of its development. It was Tycho Brache in the late 1500s who introduced the use of means as central tendencies to astronomy (Plackett, 1970). He used them to address a problem that had long troubled astronomers: what to take as the position of a star, given that the observed coordinates at a particular time tended to vary from observation to observation. When early astronomers began computing means of observations, they were very cautious, if not suspicious, about whether and when it made sense to average observations. In fact, before the middle of the eighteenth century, they would never combine their own observations with those obtained from another astronomer. They were fearful that if they combined data that had anything but very small errors, the process of averaging would multiply rather than reduce the effect of those errors (Stigler, 1986, p. 4). Taking the mean of multiple observations became the standard solution only after it had been determined that the mean tended to stabilize on a particular value as the number of observations increased.

It was another hundred years before Quetelet began applying measures of central tendency to social and human phenomena (Quetelet, 1842). The idea of applying

means to such situations was inspired partly by the surprising observation that national rates of birth, marriage, and suicides—events that at one level were subject to human choice—remained relatively stable from year to year. Some, including Arbuthnot and De Moivre, had taken these stable rates as evidence of supernatural design. Quetelet explained them by seeing collections of individual behaviors or events as analogous to repeated observations. Thus, he regarded observing the weights of 1000 different men—weights that varied from man to man—as analogous to weighing the same man 1000 times, with the observed weight varying from trial to trial. The legitimacy of such an analogy, of course, has been a heated controversy in statistics. Even at the time, Quetelet's ideas brought stiff rebukes from thinkers such as Auguste Comte, who thought it ludicrous to believe that we could rise above our ignorance of values of individual cases simply by averaging many of them (Stigler, 1986, p. 194). To Comte, statistics applied to social phenomena was computational mysticism.

We think that the way these early thinkers reacted to different applications of the mean is not merely a historical accident but instead says something about the "deep structure" of these different applications. To explore the challenges of learning to think about data as signal and noise, we examine the metaphor in the context of three types of statistical processes: repeated measures, measuring individuals, and dichotomous events.

Repeated Measures

Consider weighing a gold nugget 100 times on a pan balance, a prototypical example of repeated measurement. It almost goes without saying that the purpose of weighing the nugget is to determine its weight. But how does one deal with the fact that the observed weight varies from trial to trial? We assume that statisticians and nonstatisticians alike would regard these fluctuations as resulting from errors in the measurement process. But given this variation, how should we use the 100 measurements to arrive at the object's weight? Should all the measurements be used? Perhaps not, if they are all not equally accurate. A novice might attempt to deal with this by trying to separate the 100 measurements into two classes: those that are truly accurate versus those that are not. The problem then becomes how to tell which observations are truly accurate, because you do not know what the actual weight is.

One aspect of this situation that makes using a mean of the observations particularly compelling is that, conceptually, we can separate the signal from the noise. Because we regard an object as having some unknown but precise weight, it is not a conceptual leap to associate the mean of several weighings with this actual weight, while attributing the trial-by-trial variations to a distinctly different thing: chance error produced by inaccuracies of the measurement instrument and by the process of reading values from it. Indeed, we can also regard each individual weighing as having two components—a fixed component determined by the actual weight of the nugget and a variable component attributable to the imperfect measurement process.

The relative clarity of this example hinges on our perception that the weight of the nugget is a real property of the nugget. A few philosophers might regard it (possibly along with the nugget itself) as a convenient fiction. But to most of us, the weight is something real that the mean weight is approximating closely and that individual weighings are approximating somewhat less closely. Another reason that the idea of central tendency is compelling in repeated measurement situations is that we can easily relate the mean to the individual observations as well. To help clarify why this is so, we will make a number of assumptions explicit.

We have been assuming that the person doing the weighing is careful and that the scale is unbiased and reasonably accurate. Given these assumptions, we expect that the variability of the weighings would be small and that the frequency histogram of observations would be single-peaked and approximately symmetric. If instead we knew that the person had placed the nugget on different parts of the balance pan, read the dial from different angles, or made errors in transcribing the observations, we would be reluctant to treat the mean of these numbers as a central tendency of the process. We would also be hesitant to accept the mean as a central tendency if the standard deviation was extremely large or if the histogram of weights was bimodal. In the ideal case, most observations would be *close* to the mean or median and the distribution would peak at the average, a fact that would be more apparent with a larger data set as the histogram would be *smoother*. In this case, we could easily interpret the sample average as a good approximation to a signal or a central tendency and view the variability around it as the result of *random error*.

These assumptions about the procedure and the resulting data may be critical to accepting the mean of the weighings as a central tendency, but they are not the only things that make that interpretation compelling. As indicated above, we maintain that the key reason why the mean observation in this example is relatively easy to accept as a central tendency is that we can view it as representing a property of the object while viewing the variability as a property of a distinctly independent measurement process. That interpretation is much harder to hold when, rather than repeatedly measuring an attribute of a single object, we measure an attribute of many different objects, taking one measurement for each object and averaging them.

Measuring Individuals

Consider taking the height of 100 randomly chosen adult men in the United States. Is the mean or median of these observations a central tendency? If so, what does it represent? Many statisticians view the mean in this case as something like the *actual* or *true* height of males in the United States (or in some subgroup). But what could a statement like that mean?

There are a few reasons that an average in this situation is harder to view as a central tendency than the average in the repeated measurement example. First, the gold nugget and its mass are both perceivable. We can see and heft the nugget. In contrast, the population of men and their average height are not things we can perceive as directly. Second, it is clear why we might want to know the weight of

the nugget. But why would we want to know the average height of a population of men? Third, the average height may not remain fixed over time, because of factors such as demographic changes or changes in diet. Finally, and perhaps most importantly, we cannot easily compartmentalize the height measurements into signal and noise. It seems like a conceptual leap to regard each individual height as partly *true height*, somehow determined from the average of the population, and partly *random error* determined from some independent source other than measurement error.

For all of these reasons, it is hard to think about the average height of the group of men as a central tendency. We speculate, however, that it is somewhat easier to regard *differences* between the averages of two groups of individual measurements as central tendencies. Suppose, for example, we wanted to compare the average height of U.S. men to the average height of (a) U.S. women or (b) men from Ecuador. We might interpret the difference between averages as saying something in the first case about the influence of genetics on height and in the second, about the effects of nutrition on height. When making these comparisons, we can regard the difference in averages as an indicator of the "actual effect" of gender or of nutrition, things that are easier to imagine wanting to know about even if they are difficult to observe directly.[10]

Some support for this speculation comes from Stigler (1999), who claims that Quetelet created his infamous notion of the "average man" not as a tool to describe single distributions, but as a method for comparing them: "With Quetelet, the essential idea was that of *comparison*—the entire point was that there were different average men for different groups, whether categorized by age or nationality, and it was for the study of the nature and magnitude of those differences that he had introduced the idea" (p. 61). Although we concede that the notion of a "true" or "actual" value is still a bit strained in these comparison cases, we believe that one needs some approximation to the idea of *true* value to make meaningful comparisons between two groups whose individual elements vary. To see why, let us look more closely at the comparison of men versus women.

Suppose we compute a mean or median height for a group of U.S. men and another for a group of U.S. women. Note that the act of constructing the hypothesis that gender partly determines height requires us to conceive of height as a process influenced by various factors. Furthermore, we cannot see how comparing the two groups is meaningful unless we have (a) an implicit model that gender may have a *real* genetic effect on height that is represented by the difference between the average for men and the average for women, and (b) a notion that other factors have influences on height that we will regard as *random error* when focusing on the influences of gender on height.[11] Thus, we claim that the concept of an average

---

[10] Of course, both differences may reflect both *nature* and *nurture*.

[11] It is possible that genetic differences may also (or instead) be reflected by differences in variability in the groups. Thinking about such differences, however, also requires thinking about some sort of measure (e.g., the standard deviation or the interquartile range) as a signal reflecting the typical variability in a group.

as approximating a signal, or true value, comes more clearly into focus when we are considering the influence of a particular variable on something (in this case, gender on height). Such a comparison scheme provides a conceptual lever for thinking about signal (gender influences) and noise (other influences). We return to this point later.

Discrete Events

Another measure that is often used as an index of central tendency is the rate of occurrence of some event. As a prototypical example, consider the rate of contracting polio for children inoculated with the Salk vaccine. Even though individual children either get the disease or do not, the rate tells us something about the ability of inoculated children, as a group, to fight the disease.

How can we view a rate (or probability) as a measure of central tendency? First, a probability can be formally viewed as a mean through what some would regard as a bit of trickery. If we code the event "polio" as a 1, and the event "no polio" as a 0, then the probability of getting polio is merely the mean of these Boolean values. Producing a formal average, however, does not automatically give us a measure of central tendency. We need to be able to interpret this average as a signal related to the causes of polio. Compare the distribution of values in the dichotomous case to the ideal case of the weighing example. In the dichotomous case, the mean is not a value that can actually occur in a single trial. Rather than being located at either of the peaks in the distribution, the mean is located in the valley between, typically quite *far* from the observed values. Thus, it is close to impossible to think about the rate or probability as the *true-value* component of any single observation and the occurrence or nonoccurrence of an individual case of polio as the sum of a true value and a *random error* component. We suspect that this is largely why the idea of a central tendency in dichotomous situations is the least tangible of all.

It might help in reasoning about this situation to conceive of some process about which the rate or probability informs us. In the disease example, the conception is fairly similar to the earlier height example: A multitude of factors influence the propensity of individuals to get polio—level of public health, prior development of antibodies, incident rate of polio, age—all of which lead to a rate of getting the disease in some population. So even though individuals either get polio or do not, the propensity of a certain group of people to get polio is a probability between 0 and 1. That value is a general indicator of the confluence of polio-related factors present in that group.

As with our height example, although an absolute rate may have some meaning, we think it is much easier to conceptualize the meaning of a signal when we are comparing two rates. In the polio example, this might involve comparing the rate in an inoculated group to the rate in a placebo control group. Here, as with the height example, most people would consider the difference in rates (or the ratio of the rates) to be a valid measure of the efficacy of the vaccine or as a reasonable way to compare the efficacy of two different vaccines.

The Role of Noise in Perceiving a Collection as a Group

We have argued that the idea of central tendency, or data as signal and noise, is more easily applied to some types of processes than to others. But there are other factors, to which we have alluded, that may affect the difficulty of applying this idea. Consider the case of comparing the heights of men and women. We would expect that the shape and the relative spread of the distributions would affect how easy it is to conceive of each distribution as a coherent group and, as a consequence, to be able to interpret each group's average as an indicator of a relatively stable group characteristic.

Indeed, perhaps the most critical factor in perceiving a collection of individual measurements as a group is the nature of the variability within a group and how it relates to the differences between groups. In general, we expect that these individual measurements are easier to view as belonging to a group (and thus as having a central tendency) when the variability among them is relatively small. To explain what we mean by *relatively small*, we find the idea of *natural kinds* helpful. According to Rosch and Mervis (1975), people often mentally represent real-world concepts as prototypes and judge particular instances as "good" or "bad" depending on how closely those instances match the category prototype. For example, a prototypical bird for most North Americans is a medium-sized songbird, something like a robin. The closer an instance is to the category prototype, the less time it takes to identify that instance as a member of the category. North Americans can categorize a picture of a starling as a bird faster than they can a picture of an ostrich.

In this theory of natural kinds, prototypes function much as averages do: Instances of the category are single observations that can be some distance from the average (or prototype). In fact, some competing theories of natural kinds (e.g., Medin & Schaffer, 1978) claim that there is no actual instance that functions as a prototype, but that the effective prototype is simply a mean (in some multidimensional feature space) of all the instances in memory. What makes some categories, such as birds, natural kinds is that there is little variability across features within the category relative to the variability of those features between various animal categories. So, even though there are some nonprototypical instances of birds, such as penguins and ostriches, the distributions of features of birds overlap little with those of other natural kinds such as mammals, so that the groups cohere. This research suggests that it might be easier to accept, for example, the mean heights of the men and women as representing group properties if there were no overlap in heights of the men and women, or if at least the overlap were small relative to the spread of the distributions.[12]

Applying Central Tendency to Nonstandard Cases

In the foregoing examples, we focused on relatively ideal cases. We tacitly assumed that our histograms of people's heights, for example, were single-peaked,

---

[12] However, we should note that in the Bright and Friel (1998) study cited earlier, the two distributions were non-overlapping, yet students did not use averages to compare them.

approximately symmetric, and, configured as two histograms, had approximately equal spread. In such cases, most experts would accept some average as a meaningful measure of central tendency. Is the idea of central tendency applicable only to these ideal cases, or is it more generalizable than that? In this section, we consider several nonstandard examples to make the case that we can and do apply the idea of central tendency to less ideal situations, in which there is some doubt about whether a single measure of center is adequate to describe the data. We argue that statistical reasoning in these situations still rests to a large extent either on the conception of an average as a central tendency or on its cousin, a single measure that describes the variability of a group of observations.

Distributions with outliers. Consider cases where there are outliers that we decide should be removed from the data set. In the case of weighing, suppose a typical observation differs from the mean weight by something like 1 mg. If one of our observations was 5 mg away from the mean, most people might think it sensible to omit that value in calculating the mean. There are two ideas that seem implicit in this thinking: (a) that there is "true" measurement error associated with weighing on that scale and (b) that some *different* process can sometimes generate observations with unusually high measurement error. Only with such an implicit model can we consider, let alone decide, that an extremely deviant observation must have been due to nonrandom error (e.g., misrecording the observation or having a finger on the pan). Similarly, if we had one or two height observations that were 60 cm from the mean, we might disregard them in certain analyses as resulting from a process different from the process producing the rest of the data (e.g., from a mutation or birth defect). Here again, this makes sense only if we have some implicit model of a *typical* (male or female) height from which individual observations differ by something like "random genetic and/or environmental variation." We can then regard extremely tall or short people as not fitting this model—as resulting from a somewhat different process and therefore calling for a different explanation. For these same reasons, Biehler (1994, p. 32) suggested that "symmetrical unimodal distributions are something distinctive," and deviations from them require additional modeling.

Distributions with unusual shape. Continuing with the example of men's heights, consider the case perhaps furthest from the ideal, where the histogram of men's heights is bimodal. We would be reluctant in this case to interpret any average as a central tendency of men's heights. Why? With a bimodal histogram, we would be doubtful that the men we were looking at comprised a simple process, or "natural kind." Rather, we would suspect that our batch of men consisted of two distinct groups and that we could not make any useful statements unless we uncovered some underlying variable that distinguished the two. A similar but somewhat less severe problem would result if the histogram was unimodal but the variability in the group seemed enormous (e.g., if men's heights from an unknown country varied from 60 cm to 900 cm with a mean of 450 cm). Given the huge variability in this case, we would question whether the data came from a coherent process and

whether it made sense, therefore, to use an average to represent it. Of course, people's intuitions about whether variability is enormous may differ and are likely to depend on the model they have of *typical* variability (or indeed whether they have any conceptual model for thinking about sources of variability).

Comparing groups with skewed or differently shaped distributions. When comparing two histograms, say of men's and women's heights, we run into difficulties when the histograms are of different shape. Imagine, for example, that the men's heights were positively skewed and the women's heights negatively skewed. Because there is clearly something different about the variability in each group, we would be reluctant to compare the two groups using their averages. That is, unless we could generate a model of why the histograms of the groups differed in shape and, as a result, conclude that the different shapes were just two versions of random error, we would probably be wary of viewing the difference between the two averages as representing something like the "gender effect on height."

Consider the comparison of differences in income from one decade to another, where both histograms are highly skewed with long tails out to the right. If the histograms have the same variance and the same shape, we claim it is not unreasonable to accept the shift in central tendency as an estimate of the *actual* change in income for the group, even though we might have misgivings about using the average for either group as the best measure of *actual* income. That is, even though the variability in each group may not match our ideal view of "noise," we can at least convince ourselves that it is the same noise process in both groups. Of course, even though one histogram is a horizontal translation of the other, it does not necessarily mean that income has improved the same amount for each individual (or each type of individual), give or take random error. Indeed, a finer analysis could indicate that certain groups have become better off while other groups have not changed or have even become worse off. It is worth noting, however, that many such arguments about why looking at the differences between group averages is inappropriate or misleading rely on the perception that the groups are, in some sense, not "natural kinds" (e.g., that the processes that determine incomes of poor people are different from those that determine incomes of rich people). Nonetheless, these arguments are usually most compelling when we can identify natural subgroups in the larger group and can show that the changes in the averages in these subgroups differ from each other (e.g., the rich got richer and the poor got poorer, or different things happened to Blacks and Whites).

Another classic difficulty involves comparing two averages when the distributions differ in spread. For example, what if Country A not only has a higher mean income than Country B but also has a higher standard deviation? This would call for more serious modeling of the variability. A special case that would make it conceptually easier to compare the averages of the two groups would be the situation in which the difference in standard deviations was commensurate with the difference in means (ideally, the ratio of standard deviations would be equal to the ratio of the means). In this case, we could view the effect as multiplicative rather

than additive, since Country A's typical income would be equal to Country B's multiplied by a factor that represents the effect(s) that distinguish A from B. And it would not be unreasonable to assume that the same multiplicative factor also applied to the noise process.

Summary of analyses of nonstandard cases. As we have implied in our argument above, we do not *necessarily* see these nonstandard cases as problems for the type of framework that we are advocating. Indeed, we think that the idea of central tendency of a process allows us to (a) decide to eliminate an outlier or break data into suitable subsets, (b) come up with a conceptual model that explains why the groups are asymmetric or differ in spread or shape, or (c) decide that there is little we can sensibly conclude about the differences between the two sets of data.

Let us summarize by asking what we could conclude about the difference in men's and women's heights from the distributions we described earlier that were skewed in opposite directions. We assert that we could conclude nothing without some conceptual model. If we were trying to make a statement about genetic gender differences, for example, we would have to be convinced that everything else was *random* and that, for instance, we could not explain the mean height difference as resulting from gender differences in diet. In other words, there is virtually nothing about analyzing data that is model-free. Some may regard this a radical proposal, but we claim that a mean or median has little heuristic value (and is likely to have little meaning or heuristic value for the student) unless we can conceive of the data coming from some coherent process that an average helps to elucidate.

IMPLICATIONS FOR STATISTICS EDUCATION

The idea of noisy processes, and the signals that we can detect in them, is at the core of statistical reasoning. Yet, current curricula do not introduce students to this idea, instruments meant to assess student reasoning about data do not include items targeting it, and statistics education researchers have not given it much attention. If our argument is valid, then critical changes are called for in education research, the formulation of education objectives, curriculum materials, teacher education, and assessment. These are tightly interrelated components of educational reform. If we fail to advance our efforts on all these fronts, we run the risk of continuing to lose the small ground gained on any one of them.

Accordingly, we describe below what we see as essential components of a signal-versus-noise perspective and offer suggestions about how we might help students (and future teachers) develop these ideas. We do not aim our speculations at curriculum designers or teachers in the hope that that they will implement them. Instead, we intend them for researchers and others who are considering what big ideas should guide our standards and curriculum objectives, for those designing and running teacher institutes, and for those developing assessment frameworks and instruments.

Using Repeated Measures

According to our analysis, processes involving repeated measures are easier than other types of statistical processes to view as part signal and part noise. This suggests that to establish the signal-versus-noise interpretation of various statistical measures, we initially involve students in investigations of repeated measures.

Current curricula make little use of repeated measures. Perhaps this is because many of the prototypical situations, such as our weighing example, can be somewhat boring and seemingly pointless unless they are introduced in meaningful ways. There are many suitable and potentially interesting contexts.[13] In the later grades, these include a number of high-stakes scientific and political issues. For informed public policy, we need good estimates of the thickness of the ozone layer, of dissolved oxygen in rivers, of concentrations of atmospheric CO_2. Statistical control of manufacturing processes provides another context in which it is relatively clear why we need to track a process by looking at its outputs. Of course, time series analyses are complex, and we need more research to help determine the kinds of questions regarding them that introductory students can fruitfully explore.

Lehrer, Schauble, and their colleagues have employed some interesting repeated measure contexts with younger students. For example, students in a second-grade class designed cars to race down a track (Lehrer, Schauble, Carpenter, & Penner, 2000). During trial runs, students became unhappy about a decision to base a claim about a car's speed on a single trial. Frequently, something would happen to impede a car—for example, it would run up against the track's railing. The agreed-on remedy was to race each car five times. Not surprisingly, the students could not agree later on how to get a single measure of speed from the five trials. However, their proposal of multiple trials was, by itself, suggestive of some notion of signal (a car's actual top speed on that track) and noise (its observed times resulting from unpredictable events).

This classroom episode suggests an important distinction. That is, a student might perceive data as comprising signal and noise and yet not necessarily view a statistical measure such as an average as an acceptable indicator of signal. We would expect that with processes involving repeated measures, students would tend to think of each measurement as a combination of signal and noise, particularly if sources of measurement error were easy to identify, as in measuring length with a ruler. But these same students might not be likely to think of an average of repeated measures as indicative of signal (any more than the early astronomers were). Thus, the instructional challenge is how to help students interpret measures such as averages as indicators of central tendency. Taking a clue from the historical development of the concept, it would seem fruitful to have students explore the relative stability of various indicators in different samples.

---

[13] For several good examples of activities written around such processes, see Erickson (2000).

Explorations of Stability

The idea of stability is closely related to the idea of signal. If the weight of an object is not changing from trial to trial, it seems reasonable to expect that a good indicator of its weight should also not vary much from sample to sample. Recall that it was observing the stability from year to year of such things as birth and death rates that led Quetelet to begin regarding these rates as indicators of prevailing and relatively stable societal conditions, and to make the analogy to means of repeated measures. Similar investigations by students could set the stage for interpreting averages as indicators of signal.

A method frequently used to demonstrate stability is to draw multiple samples from a known population and evaluate particular features, such as the mean, across these replications. However, we expect that these demonstrations are often conducted prematurely—before students have understood why one is interested in the mean. Furthermore, in real sampling situations we never do these repeated samplings, which leaves many students confused about what we can possibly learn from this hypothetical exercise. The following three alternative methods of exploring stability appear promising on the basis of their use in classrooms with students as young as 8 years old.

Comparing different measures. In this approach, students compare the relative accuracy of different measurement methods. Lehrer, Schauble, Strom, and Pligge (2001) used this approach with third and fifth graders, who measured weights and volumes as part of a study of densities of different materials. The students explored a number of different ways to measure each attribute. They did this by using each method repeatedly to measure the same object. The students came to favor those methods that produced less variability in these repeated measures. Having established what measurement technique they would use, they then considered various proposals of what to use as, for example, the volume of a particular object. The problem, of course, was that even with the same measurement method, repeated measuring gave the students a range of values. They ultimately decided to discard outliers and compute the means of the remaining observations as their "best guess" of the weights and volumes of these objects.

Observing growing samples. Another way of exploring stability is to have students observe a distribution as the sample gets larger. We tested this approach recently in a seventh-grade mathematics class. Students had conducted an in-class survey to explore whether boys and girls were paid similar allowances. While comparing the two distributions, one student expressed reservations about drawing conclusions, arguing that she had no idea what the distributions might look like if they collected more data. Her classmates agreed.

To help the class explore this issue, we constructed an artificial pond filled with two kinds (colors) of paper fish. According to our cover story, a farmer wanted to determine whether a new type of genetically engineered fish grew longer, as claimed, than the normal fish he had been using. Students "captured" fish from

the pond, reading off fish type and length (which was written on the fish.) On an overhead display, we constructed separate stacked dot plots for each type of fish as students read off their data. After about 15 fish had been sampled, we asked students what the data showed so far. Students observed that the data for the normal fish were clustering between 21–24 cm, whereas the data for the genetically engineered fish were clustering between 25–27 cm. Then we asked them what they thought would happen as we continued to sample more fish, reminding them of their earlier reservations with the allowance data. Some said that the stacks would become higher and the range would get bigger, without mentioning what would happen to such features as the general shape or the location of the center clump. However, other students did anticipate that the center clusters would "grow up" but would nevertheless maintain their approximate locations along the horizontal axis. The latter, of course, is what they observed as they continued to add more fish to the sample distributions. After the sampling, we showed them both population distributions along with their sample data, calling their attention to the fact that the centers of their sample distributions were quite good predictors of the centers of the population distributions—that these stable features of the samples were signals.

Simulating processes. A third way to explore stability is to investigate why many noisy processes tend to produce mound-shaped distributions. Wilensky (1997) described a series of interviews that he conducted with graduate students who were exploring this question through computer simulations. We conducted a similar investigation with fifth-grade students in an after-school program on data analysis. In analyzing a data set on cats (from Rubin, Mokros, & Friel, 1996), students noticed that many frequency distributions, like tail length and body weight, were mound shaped. As part of exploring why this might be, students developed a list of factors that might cause a cat's tail to be longer or shorter. Their list included diet, being in an accident, and length of father's and mother's tails. Using this list, we constructed a spinner to determine the value of each factor for a particular cat's tail. One student might spin +2 inches for diet, +3 inches for mother's contribution, -2 inches for an accident, etc. (Of course, each student wanted his or her cat to have the longest tail.) Before they began spinning, students predicted that if they built 30 cat tails in this way, they would get about equal numbers of cats with short, medium, and long tails. After several trials they noticed they were tending to get medium tails, which they explained by pointing out that you would have to be "real lucky" to get a big number every spin, or "real unlucky" to get a small number every spin. As this was our last session with these students, we could not explore what they might have generalized from this experience, but we believe that understanding why such processes produce normal-shaped distributions is a critical part of coming to trust how process signals rise up through the noise.

Group Comparison

We have speculated that it is often easier to regard the difference between two averages as a central tendency than it is to think of a single average that way. This suggests, perhaps somewhat counterintuitively, that rather than beginning instruction by having students explore single distributions of individual values, we instead might fruitfully start with questions involving group comparison. Some support for the benefit of having even young students grapple with comparison problems comes from accounts from teachers of data analysis in the elementary grades (Konold & Higgins, in press). Similarly, all the problems in the middle-school materials developed by Cobb, McClain, and Gravemeijer involve group comparison (Cobb, 1999; Cobb, McClain, & Gravemeijer, in press). As Watson and Moritz (1999) pointed out, some of the benefits of comparison contexts are undoubtedly related to their being more interesting and allowing students to see more clearly why the question matters and why averages might be useful. But in addition, we expect that in a comparison situation, students can more easily view averages of the individual groups as summary measures of processes and can readily perceive the difference between those measures as some signal rising through the din of variability.

Conducting Experiments

Many educators have touted the benefits of students' collecting their own data (e.g., Cobb, 1993). Among the expected advantages are increased student interest and the rich source of information that students can draw on as they later analyze and reason about the data. There may be additional benefits to having students design and run simple, controlled experiments. One benefit derives from the fact that experimental setups involve group comparison. In addition, we speculate that data from experiments are easier than observational data to view as coming from a process. As experimenters, students take an active role in the process—for example, by fertilizing one group of plants and comparing their growth to that of an unfertilized group of plants. Even quite young students can understand the importance in such cases of treating both groups of plants the same in all other respects (Lehrer, Carpenter, Schauble, & Putz, 2000; Warren, Ballenger, Ogonowski, Rosebery, & Hudicourt-Barnes, 2001). They then observe firsthand that not every plant in the fertilized group responds the same and that the effect of the fertilizer becomes evident, if at all, only when comparing the two groups. With observational data, students must reason backwards from observed differences to possible explanations for those differences, and their tendency in explaining the data is to offer different causal accounts for each individual value. With the experimental setup, students first see the process and then the data resulting from it, a difference in perspective that may help them focus on the class of causes that apply uniformly at the group, as opposed to the individual, level.

CONCLUSIONS

We fear that some readers will hear in our analysis and recommendations a call to abandon the teaching of noninferential exploratory methods of data analysis and to eschew data from other than well-defined samples. In fact, we believe that we should begin teaching informal methods of data analysis in the spirit of EDA to students at a young age. Moreover, we are not recommending that the teaching of data analysis should be grounded in, or necessarily headed towards, the technical question of drawing formal inferences from carefully constructed samples.

We agree with Tukey (1977) that we should not, as a rule, approach data with the knee-jerk desire to model them mathematically. Rather, our objective should be more general—to learn from them. For this purpose, being able to display data flexibly and in various ways can lead to interesting insights and hypotheses, some of which we may then choose to model more formally (Cleveland, 1993). It is this sensible approach to the general enterprise—not only to *how* but also to *why* we collect and explore data—that we believe is most important to convey to students in early introductions to statistics.

It is important that we keep in mind, however, that most of us who regularly use exploratory methods of data analysis have strong backgrounds in inferential methods. When we approach data exploration with fewer assumptions, we often set aside, for the moment, much of the power of the mathematical models of statistics. But to play data detective, we have a host of tools and experiences to draw on, many of which stem from our knowledge of the mathematical models of statistics. As Cleveland (1993) observes, "Tools matter (p. 1)." The tools that he was referring to were methods of displaying data. We would add that underlying the skillful use of such graphical tools is the skillful use of conceptual ones, which matter even more.

Our references to the pioneering work of Quetelet were meant to highlight the fact that the early users of means did not regard them simply as ways to describe centers of distributions, which is how some today (misleadingly) characterize them. Recent histories of the development of statistics (Hacking, 1990; Porter, 1986; Stigler, 1986) portray the early innovators of statistics as struggling from the beginning with issues of interpretation. In this regard, Quetelet's idea of the "average man" was a way to take the interpretation of a mean as a "true value" of repeated measures and bootstrap it to a new domain—measurements of individuals—for which the mean did not initially make much intuitive sense. We believe that learning to reason about data requires students to grapple with the same sorts of interpretation issues, and that in the process they need to develop conceptual (not necessarily mathematical) models of data that can guide their explorations. The idea of data as signal and noise, physically embodied in the workings of the Galton Board (see Biehler, 1994), is perhaps the most fundamental conceptual model for reasoning statistically. Future research should help us learn how the idea develops and how we can foster that development in our students.

REFERENCES

American Association for the Advancement of Science (AAAS). (1989). *Science for all Americans.* Washington, D.C.: AAAS.

Bakker, A. (2001). Historical and didactical phenomenology of the average values. In P. Radelet-de Grave (Ed.), Proceedings of the Conference on History and Epistemology in Mathematical Education (Vol. 1, pp. 91–106). Louvain-la-Neuve and Leuven, Belgium: Catholic Universities of Louvain-la-Neuve and Leuven.

Biehler, R. (1989). Educational perspectives on exploratory data analysis. In R. Morris (Ed.), *Studies in Mathematics Education* (Vol. 7, pp. 185–201). Paris: UNESCO.

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.

Bright, G. W., & Friel, S. N. (1998). Helping students interpret data. In Lajoie, S. P. (Ed.), *Reflections on statistics: Learning, teaching, and assessment in grades K-12* (pp. 63–88). Mahwah, NJ: Lawrence Erlbaum Associates.

Cleveland, W. S. (1993). *Visualizing data.* Summit, NJ: Hobart Press.

Cobb, G. (1993). Reconsidering statistics education: A National Science Foundation conference [Electronic version]. *Journal of Statistics Education, 1*(1), Article 02.

Cobb, P. (1999). Individual and collective mathematical development: The case of statistical data analysis. *Mathematical Thinking and Learning, 1*(1), 5–43.

Cobb, P., McClain, K., & Gravemeijer, K. (in press). Learning about statistical covariation. *Cognition and Instruction.*

Cortina, J., Saldanha, L., & Thompson, P. (1999). Multiplicative conceptions of the arithmetic mean. In F. Hitt & M. Santos (Eds.), *Proceedings of the 21st Meeting of the North American Chapter of the International Group of the Psychology of Mathematics Education,* (Vol. 2, pp. 466–472), Cuernavaca, Mexico: Centro de Investigación y de Estudios Avanzados.

Donahue, P. L., Voelkl, K. E., Campbell, J. R., & Mazzeo J. (1999) *NAEP 1998 reading report card for the Nation and the States.* Document No. NCES 1999-500. Washington, DC: National Center for Educational Statistics, U.S. Department of Education. Available from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=1999500

Erickson, T. (2000). *Data in depth: Exploring mathematics with Fathom.* Emeryville, CA: Key Curriculum Press.

Feldman, A., Konold, C., & Coulter, R. (2000). *Network science, a decade later: The Internet and class-room learning.* Mahwah, NJ: Lawrence Erlbaum Associates.

Freund, R. J., & Wilson, W. J. (1997). *Statistical methods.* Boston: Academic Press.

Frick, R. W. (1998). Interpreting statistical testing: Process and propensity, not population and random sampling. *Behavior Research Methods, Instruments, & Computers, 30*(3), 527–535.

Gal, I., Rothschild, K., & Wagner, D. A. (1990). *Statistical concepts and statistical reasoning in school children: Convergence or divergence.* Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Gordon F. S., & Gordon S. P. (1992). *Statistics for the twenty-first century.* (MAA Notes, no. 26). Washington, D.C.: Mathematical Association of America.

Gould, S. J. (1996). *Full house.* New York: Harmony Books.

Hacking, I. (1990). *The taming of chance.* Cambridge, UK: 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.

Konold, C. (2002). Teaching concepts rather than conventions. *New England Journal of Mathematics, 34*(3), 69–81.

Konold, C., & Garfield, J. (1992). *Statistical Reasoning Assessment: Intuitive Thinking*. Unpublished Manuscript. Amherst, MA: University of Massachusetts.

Konold, C., & Higgins, T. (in press). Reasoning about data. In J. Kilpatrick, W. G. Martin, & D. E. Schifter (Eds.), *A research companion to principles and standards for school mathematics*. Reston, VA: NCTM.

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.

Konold, C., Robinson, A., Khalil, K., Pollatsek, A., Well, A., Wing, R., & Mayr, S. (2002). Students' use of modal clumps to summarize data. Paper presented at the Sixth International Conference on Teaching Statistics, Cape Town, South Africa.

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

Lehrer, R., Schauble, L., Carpenter, S., & Penner, D. (2000). The inter-related development of inscriptions and conceptual understanding. In P. Cobb, E. Yackel, & K. McClain (Eds.), *Symbolizing and communicating in mathematics classrooms: Perspectives on discourse, tools, and instructional design* (pp. 325–360). Mahwah, NJ: Lawrence Erlbaum Associates.

Lehrer, R., Schauble, L., Strom, D., & Pligge, M. (2001). Similarity of form and substance: Modeling material kind. In D. Klahr & S. Carver (Eds.), *Cognition and instruction: 25 years of progress.* (pp. 39–74). Mahwah, NJ: Lawrence Erlbaum Associates.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238.

Mokros, J., & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education, 26,* 20–39.

Moore, D. S. (1990). Uncertainty. In L. A. Steen, (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95–137). Washington, DC: National Academy Press.

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

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

National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.

Noss, R., Pozzi, S., & Hoyles, C. (1999). Touching epistemologies: Meanings of average and variation in nursing practice. *Educational Studies in Mathematics, 40,* 25–51.

Plackett, R. L. (1970). The principle of the arithmetic mean. In E. S. Pearson and M. G. Kendall (Eds.), *Studies in the history of statistics and probability* (pp. 121–126). London: Charles Griffen & Company.

Pollatsek, A., Lima, S., & Well, A. (1981). Concept or computation: Students' misconceptions of the mean. *Educational Studies in Mathematics, 12,* 191–204.

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

Quetelet, M. A. (1842). *A treatise on man and the development of his faculties*. Edinburgh, Scotland: William and Robert Chambers.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 8,* 382–439.

Rubin, A., Mokros, J., & Friel S. (1996). *Data: Kids, cats, and ads. Investigations in number, data, and space*. Palo Alto, CA: Dale Seymour Publications.

Scheaffer, R. (1991). The ASA-NCTM Quantitative Literacy Program: An overview. In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics* (pp. 45–49). Voorburg, The Netherlands: International Statistical Institute Publications.

Schwartzman, S. (1994). *The words of mathematics: An etymological dictionary of math terms used in English*. Washington, DC: Mathematical Association of America.

Shaughnessy, J. M., Watson, J., Moritz, J., & Reading, C. (1999). School mathematics students' acknowledgment of statistical variation. Paper presented at the 77th annual meeting of the National Council of Teachers of Mathematics, San Francisco.

Smith, G. (1998). Learning statistics by doing statistics [Electronic version]. *Journal of Statistics Education, 6*(3), Article 04.

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

Stigler, S. M. (1999). *Statistics on the table: The history of statistical concepts and methods.* Cambridge, MA: Harvard University Press.

Strauss, S., & Bichler, E. (1988). The development of children's concepts of the arithmetic average. *Journal for Research in Mathematics Education, 19*, 64–80.

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

Warren, B., Ballenger, C., Ogonowski, M., Rosebery, A., & Hudicourt-Barnes, J. (2001). Rethinking diversity in learning science: The logic of everyday languages. *Journal of Research in Science Teaching, 38*, 1–24.

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. (2000). The longitudinal development of understanding of average. *Mathematical Thinking and Learning, 2*(1), 9–48.

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

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

Authors

Clifford Konold, Scientific Reasoning Research Institute, 424 Lederle Graduate Research Tower, University of Massachusetts, Amherst, MA 01003-9337; konold@srri.umass.edu

Alexander Pollatsek, Department of Psychology, Tobin Hall, University of Massachusetts, Amherst, MA 01003-7710; pollatsek@psych.umass.edu