CHAPTER TWO

Literature Review

Introduction

The purpose of this chapter is to present a selected review of the research on stochastical teaching and learning that is germane to the study of variation. By situating the study within the field of stochastics education, we can see how previous research in probability and statistics relates to the specific issue of knowledge about variation.

In structuring this literature review, first some studies that look at foundational aspects of probabilistic thinking are discussed. Next, examples of research on judgment heuristics, biases, and other stochastical intuitions and misconceptions are provided. Then, some findings on graphicacy and averages that focus on data handling are explicated. Lastly, some emergent research on students' concepts of variation in data sets and in sampling and probability contexts is presented. Although this literature review focuses directly on issues of variation, connections are drawn throughout the chapter on how other research relates to and informs the present study.

Some of the research has more of a probabilistic flavor and other research more statistical, but is in some sense an artificial division which separates these twin domains of stochastics. Both domains enhance and influence one another. The philosophical approach of this study is aligned with "forging connections between the study of data analysis and probability concepts" (Shaughnessy, Garfield, and Greer, 1996, p. 206).

Foundational Aspects of Probabilistic Thinking

In choosing this topic for launching the literature review, my main motive is to focus on people's appreciation of random phenomena. The ability to discriminate between the deterministic and the random lies at the heart of stochastics. In particular, the lack of appreciation for randomness can hinder one's ability to deal appropriately with variation.

Piaget and Inhelder (1975) indicated that the age threshold for appreciating the unpredictability of random phenomena is seven years, but Kuzmak and Gelman's (1986) evidence suggested "an earlier understanding of random phenomena than previously has been reported" (p. 565). Fischbein, Pampu, and Minzat (1975) provided evidence that children as young as five have some sensitivity to uncertainty. Kuzmak and Gelman (1986) used both a random device and a deterministic device in an experiment with young children. The deterministic device consisted of a transparent tube which dispensed colored balls one at a time. The color of the next ball could readily be seen. The random device was a rotating wire cage full of colored balls which could dispense these balls one at a time through an attached cup. Children who used these devices were asked whether or they not they knew for sure which color they were going to obtain. The researchers found that children between the ages of four and seven were able to understand that one cannot say for certain what will be the outcome when dealing with random events. The lack of certainty for individual outcomes is the essential feature of randomness which distinguishes it from determinism.

Research has found that the inclination to cling to a deterministic outlook even in the face of a random process is tenacious. Working with second-graders, Horvath and Lehrer (1998) discuss how some of the children initially "did not think of rolling dice as completely random", but that "beliefs about lucky numbers and partial telekinesis usually did not endure long after the children's first opportunities to collect data" (p. 126). In a questionnaire item administered to 1014 students from grades 3, 6, and 9, 113 responses affirmed some kind of belief in lucky numbers, as exemplified by the student who wrote: "I don't think many numbers are lucky. But I think 4, 7, and 9 are, so I guess I'd agree in a way you can have lucky numbers" (Watson, Collis, and Moritz, 1995, p. 553). This attitude, similarly exemplified by the student cited by Horvath and Lehrer (1998) who said "I usually roll 6s" (p. 137), begs for an examination of what is really involved in a random event.

Green (1983), asked students aged 11 through 16 to predict the location of snowflakes falling on a 16-square grid. Choices for snowflake dispersal were given, with some choices more suggestive of a random phenomena than others. The spirit of the task is captured in Figure 1 below.

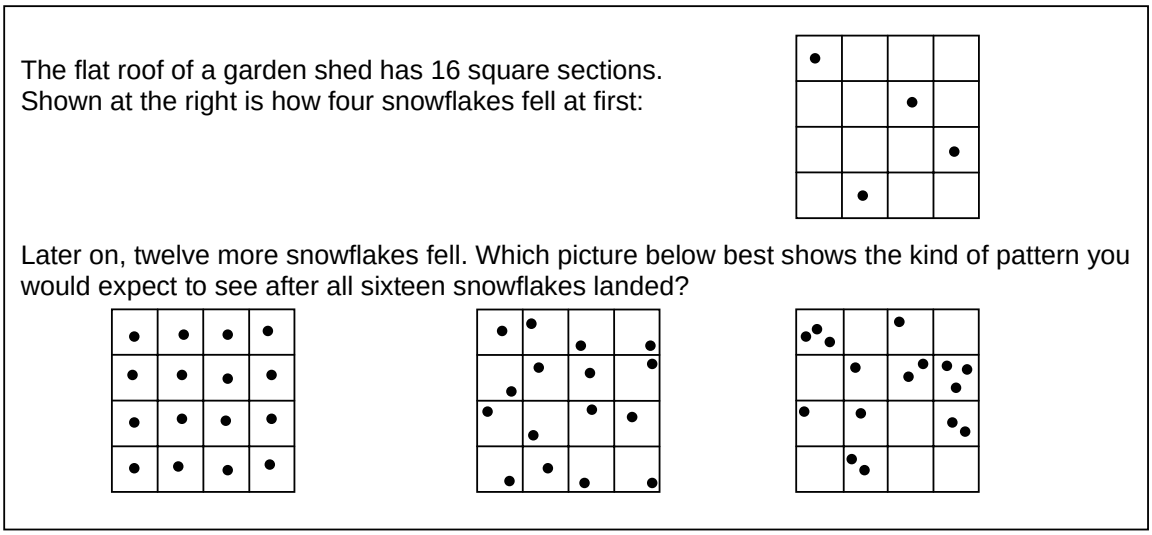

Fig. 1

Green noted, “The astonishing thing about this item is that performance declines with age” (1983, p. 774), meaning that older subjects did worse than younger subjects. Green also gave students a task in which they were presented with two sequences of coin flips. Each sequence contained 150 flips. Students were asked which of the two sequences was actually the result of carrying out the flips of a fair coin, and which sequence was simply made up in the imagination. Batanero, Green and Serrano (1998) suggest that to succeed in this task, “children need to discriminate between a random and a non-random sequence,” but that “in fact, most of the children chose the non-random sequence, and the perception of randomness did not improve with age” (p. 120). Batanero and Serrano (1999) extended Green’s results to 17-year-old students. The researchers had the following comment about their subjects:

> “On the one hand, they perceived the local variability…and unpredictability of the random processes underlying the tasks we gave them. On the other hand…students overemphasized

> unpredictability and luck to justify their attributions of randomness" (p.565).

To illustrate the meaning of this comment, first consider one of the sequences of 40 coin flips shown to the students: HTHTTHHTHTHHTTHTTHHTTHTHHTTHTHTHTHTHTTHT. This sequence has proportions of heads and tails commensurate with the theoretical probability, but has too many alterations between heads and tails. That is, the lengths of the runs are too short. Students viewing the sequence as random were thought by the researchers to have been focusing too much on the local variability inherent in the process of flipping coins.

Secondly, consider another of the sequences shown to the students: HTTTHTTHTHTTTHTTTTHHTTTHTT HTTHTTTTHTTTHT. This sequence has appropriate length of runs but a disproportionate amount of tails. Students viewing the sequence as random were considered to be placing undue emphasis on the overall unpredictability of the experiment. The range of responses of the students demonstrate the natural tension between local unpredictability versus global trends, which are key themes of randomness. The use of active experimentation, allowing for the integrated study of probability and statistics, was promoted by Batanero et. al. (1998) as a way for students to gain an understanding of the essential characteristics of random phenomena.

An appreciation for randomness cannot be overemphasized as an important facet for understanding variation, and represents something "fundamental to reasoning within the domain of statistics and probability" (Metz, 1998, p. 156). Determinism thwarts randomness, and thus stifles an appreciation of variation. Moreover, it would be an erroneous assumption to assume that by the time children (or even adults) reach a certain age, they have completely abandoned deterministic notions and embraced the truth of living in a stochasticized world.

Heuristics, Biases, Intuitions and Misconceptions

A host of misconceptions related to probabilistic and statistical thinking have been researched, and psychologists in particular have characterized many examples of faulty human

intuition in the face of uncertainty. The use of the term *misconception* in this area of research refers primarily to erroneous thinking which contrasts with "correct", normative responses. This section illustrates some of the key findings that research on intuitive stochastical thinking has produced. Because this domain of research is so robust and stable, it suggests that when subjects reason about variation, they will also bring with them strong intuitions that influence their thinking.

Psychologists Daniel Kahneman and Amos Tversky found that people come to the table of stochastical learning with a host of their own intuitions about the subject, and that these intuitions often serve them quite poorly. As they wrote early on, "We submit that people view a sample randomly drawn from a population as highly representative, that is, similar to the population in all essential characteristics" (Kahneman & Tversky, 1971, p. 105). Some examples will illustrate the key features of this judgment heuristic, which is known as *representativeness*.

For the first example, in considering the gender of six children in a family, many people consider the sequence of GGBGBB to be more likely to occur than BBBBBG. Many people think that any sample drawn from the population should reflect a near 50-50 distribution of boys and girls (Kahneman & Tversky, 1972; Shaughnessy, 1977). Kahneman and Tversky also identified what they called the "law of small numbers, which asserts that the law of large numbers applies to small numbers as well" (1971, p. 106). Shrage (1983) also observed this in his tertiary students. Using a population mixture of 50% white balls and 50% black balls, he found that students believed the chances of getting 7 white balls in 10 draws and the chances of getting 70 white balls in 100 draws were the same. People often intuitively rely on the law of small numbers when playing games of chance. If the results of a sequence stray to far from the population proportion, a "corrective bias in the other direction is expected" (Kahneman & Tversky, 1971, p. 106). This helps explain why a person who sees flips of a fair coin result in

five heads in a row will intuitively think there is a higher likelihood of a tails on the sixth flip, a psychological phenomena known as the *Gambler's Fallacy*.

The representativeness heuristic also suggests that any uncertain event must "also reflect the properties of the uncertain process by which it is generated" (Kahneman & Tversky, 1972, p. 434). For example, when flipping a coin, the representativeness heuristic implies that any sequence of flips should appear random. Thus, people do not consider sequences such as THTHTHTHTH or TTTTTHHHHH to be representative, although both have a 50-50 mix of Hs and Ts. These sequences appear to be too ordered. They do not appear to be random, and thus do not "represent the fairness of the coin" (Tversky & Kahneman, 1974, p. 1125).

A second judgment heuristic, called *availability*, is used when people ascribe likelihoods to situations based on how readily they bring similar examples to mind (Tversky & Kahneman, 1974). An example of this bias occurs when people are asked if it is easier to choose committees of two or committees of eight from a group of ten. Those operating with an availability heuristic will find it easier to think of examples of committees of two, and hence they fail to recognize the equivalent nature of the problem (Shaughnessy, 1993). Similarly, people erroneously tend to think there are more words beginning with a "r" than there are words with "r" as the third letter, because words starting with a "r" are easier to recall (Tversky & Kahneman, 1974). Reliance on the availability heuristic leads people to depend on what they can recall or mentally construct most readily.

Another facet of intuitive thinking discovered in research is the *outcome approach*, which colors a person's fundamental understanding of the goal of probability. People who operate under the outcome approach interpret probability questions as "predicting the results of a single trial" (Konold, Pollatsek, Well, Lohmeier, & Lipson, 1993, p. 394). For example, consider a six-sided die with five black sides and one white side. One student reasoned that since a single toss would "almost certainly" result in black, then by extension six tosses would result in six blacks (Konold, 1989, p. 83). Another component of the outcome approach includes the way in

which probabilities are judged as right or wrong after each outcome. For example, Konold (1989) found that, if a weather forecaster predicts a 70% chance of rain for tomorrow, and then it doesn't rain, some subjects evaluate the 70% prediction as erroneous. That is, the forecaster's prediction is judged wrong because the predicted outcome *didn't* occur. Outcome approach thinking can also result in an emphasis on causal features rather than frequency data. Some people in Konold's (1989) study inspected the features of a seven-sided irregular polyhedron, or the way in which it was rolled, to predict outcomes, and these people discounted the actual results of 1,000 previous rolls. A view towards an equiprobable interpretation may interact with outcome approach thinking. A person may view all outcomes of an event as equally likely, and deem the goal of probability questions as simply a matter of choosing any outcome which *could* occur (Shaughnessy & Ciancetta, 2001; Shaughnessy & Bergman, 1993; Shaughnessy, 2001).

Comment: This is the "White Paper"

These three types of intuitive stochastical thinking (representativeness, availability, and the outcome approach) are examples of cognitive strategies which may influence a person's reasoning in situations where attention to variation is critical. For instance, representativeness leads to a person's insensitivity to the effects of sample size. This is readily illustrated in the results shown for what is known as the Hospital Problem (Tversky & Kahneman, 1974), which is summarized as follows:

> A certain town has two hospitals. In the larger hospital, about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. For a period of 1 year, each hospital recorded the days on which more than 60% of the births were boys. Which hospital do you think recorded more such days?

The researchers noted that 53% of undergraduate students thought that the two hospitals would have recorded about the same number of such days. The subjects did not recognize that the larger hospital would be less likely to deviate from the expected 50-50 gender mix, and "This fundamental notion of statistics is evidently not part of people's repertoire of intuitions" (1974, p. 1125). Availability may lead a person to attach inflated significance to an event which has

personal meaning. Such an event becomes more than just another data point, and can bias statistical thinking. Shaughnessy (1992) offers this example, "If several of your friends have recently divorced their spouses, you may be led to believe that the local incidence of divorce is on the rise, when in fact it has not changed" (p. 472). In this sense, availability can reduce the perception of variation. If a child recalls a game of chance in which she rolled many sixes, then it may make sense for her to think that six is a lucky number. The outcome approach also can diminish the perception of variation, as demonstrated by the subjects in Konold's (1989) study who predicted all blacks in the tosses of the die with five black faces and one white face.

To summarize, when considering conceptions of variation it is important to remember the types of judgments that people are susceptible to, and that people "already have their own built-in heuristics, biases, and beliefs about probability and statistics" (Shaughnessy, 1992, p. 472). It is also important to attend to a person's fundamental understanding of randomness. The next two sections discuss research on graphicacy and averages, two aspects of data handling which both influence conceptions of variation. For instance, graphs summarize data and reveal or disguise variation in the data set depending on the type of graph used. An average might be representative of the data set, but variation can address how the data clusters or spreads out around the average.

Aspects of Data Handling: Graphicacy

Graphicacy is a term first introduced by Balchin and Coleman in 1965, and has evolved in definition until Wainer narrowed it to mean "the ability to read graphs, defining it as proficiency in understanding quantitative phenomena that are presented in a graphical way" (Friel & Bright, 1996, p. 1). The facility to construct and interpret graphs begs the question of what exactly constitutes a graph. For the purposes of this study it seems natural to consider the standard sorts of graphs and plots of "univariate data that dominate the school curriculum, that is, line plots, bar graphs, stem-and-leaf plots, and histograms" (p.1). Also, the kinds of pictographs commonly found in the media are of interest in this study, as are rudimentary presentations of

bivariate data such as scatter plots or line graphs. This study adopts the position that "the use of graphs and other kinds of representations needs to be viewed as part of the process of statistical investigation and not as an end in itself" (Friel, Bright, Frierson, & Kader, 1997, p. 62). Graphs can be viewed as an important part of data handling, whose chief role lies in data reduction.

Curcio (1987) conducted a study of 204 fourth-grade and 185 seventh-grade subjects. Students were given a test which used twelve graphs equally distributed among the following four types: bar graphs, circle graphs, line graphs, and pictographs. Curcio was able to identify three levels of difficulty in making sense of graphs, or graphicacy. The first level is simply *reading the data*, in which the subjects could attend to the basic facts stated in the graph, including numerical values shown, titles, and axis labels. For example, in a bar chart showing the heights of four children (see Fig. 2), a question invoking a simple reading of the data would be to ask, "How tall is Mark?"

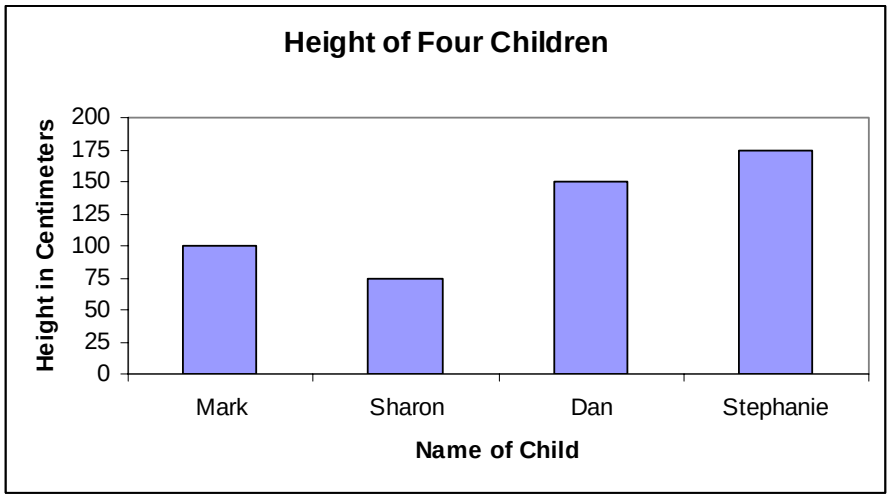

Fig. 2

The second level of comprehension is *reading between the data*, which necessitates "comparisons and the use of mathematical concepts and skills" (Curcio, 1987, p. 384). A question for this level would be, for example, "How much taller is Dan than Mark?"
The third level is *reading beyond the data*, which requires "extension, prediction, or inference" (p. 384). A question for this level would be "If Dan grows 5 centimeters and Mark grows 10 centimeters, then who will be taller, and by how much?" To establish the level of statistical literacy essential to functioning in a modern society constantly assailed by data, people need to move up through these hierarchical levels of difficulty of graphical comprehension (Moritz & Watson, 1997; Curcio, 1987).

Comment: Paragrap h on Media (see R1 p. 10 and 11:To get CUT from R2 ?

Moritz and Watson (1997) note that "reading beyond the data is necessary for questioning claims associated with graphs which are made in the media" (p. 345). The researchers presented a media-related graphical advertisement, along with the associated claim of the advertiser, to over 1800 students in grades 6, 8, 9, and 11. They found that many students did not notice difficulties in uneven scaling, or they had trouble in applying numeracy skills when interpreting data from the graph, or they could not relate the graph to the context in which it was situated. In essence, findings suggested that graphs in advertisements stand an excellent chance of confounding all levels of graphicacy. The researchers conclude that

> "Students need to be challenged in the classroom using non-standard graphs, even those with errors, to question why the author has represented a message in certain ways and to be on the lookout for misleading representations" (1997, p. 350).

The idea of assessing statistical thinking using graphs from the media was continued by Watson (1997) using 670 students in either the sixth or ninth grades. She found responses which ranged from naïve and intuitive skills to more sophisticated, interpretive skills. Using a pie chart (see Fig. 3) whose pieces were labeled with percentages that did not correspond to the percentage of area of the piece (and hence the labels did not sum to 100%), Watson found that

many students were unable to notice anything unusual at all about the graph, while others had trouble applying their knowledge to the context at hand.

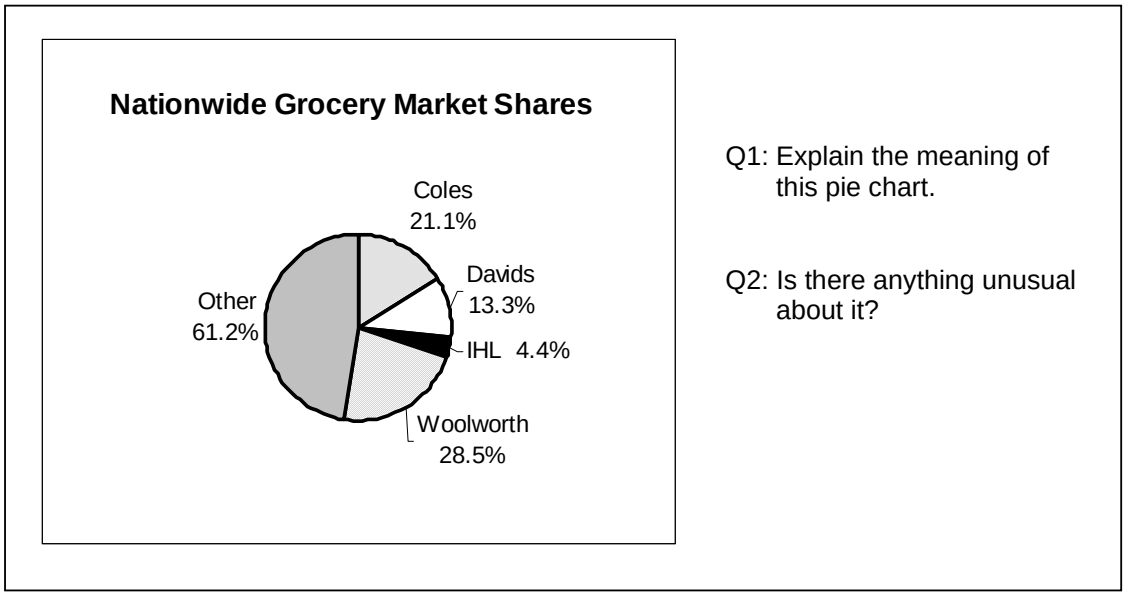

Fig. 3

Two aspects of graphicacy, the “basic understanding of statistical representation”, and the appreciation of “how that representation is embedded in a wider context” (Watson, 1997, p. 115), comprise the first two tiers of Watson’s proposed hierarchy in assessing statistical thinking. The first question in Figure 4 addressed these two tiers. The second question in Figure 4 addressed a third tier, which some students were able to achieve by questioning the percentages in the pie chart. For example, one third-tier response was, “Where it has Other, it says 61.2% and the percentage of that section on the pie is less than 50%” (p. 116).

In 1994, Friel and Bright (1996) studied graphicacy among students in grades 6, 7, and 8. They researched the levels of graph comprehension outlined earlier by Curcio (1987). For example, they showed students a line plot depicting the quantity of raisins in half-ounce boxes (see Fig. 4). Rather than initially ask students questions about *reading the data*, Friel and

Bright started with the following questions aimed at *reading between the data*: “Are there the same number of raisins in each box? How can you tell?” (Friel & Bright, 1996, p. 4).

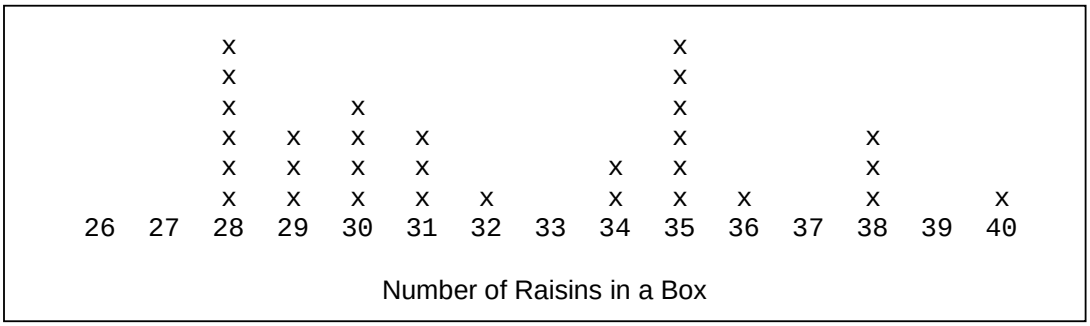

Fig. 4

Results showed that although many students were able to answer the first question correctly in the negative, in their subsequent explanations it became clear that they were looking at the graph in ways that did not support their answer. For example, one student said “No, because they weigh the boxes until they equal ½ ounce. They don’t count the raisins” (p. 5).

Bright and Friel (1998) also tested students before and after an instructional unit which was designed to “highlight connections between pairs of graphs” (p. 68). For each pair of graphs, the same set of data was used. Two pairs of graphical types were stem-and-leaf plots versus histograms and line plots versus bar graphs. A third pair was bar graphs for grouped versus ungrouped data (see Fig. 5).

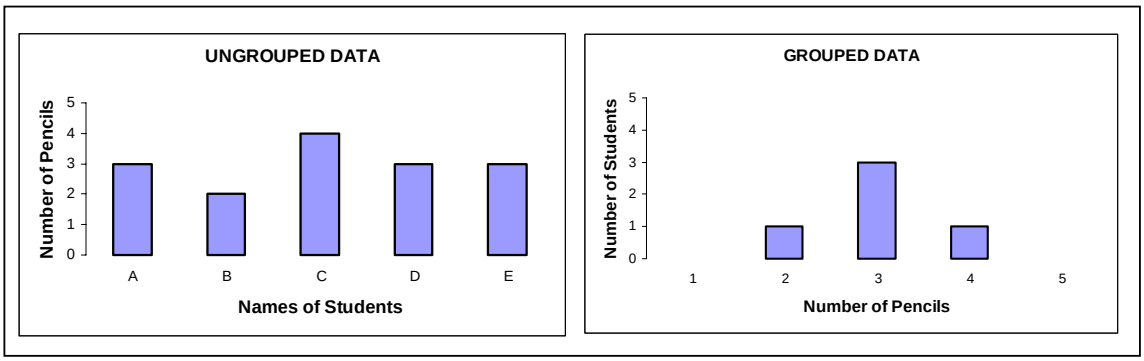

Fig. 5

The graph on the left of Figure 6 represents *ungrouped* data (raw data), and students can simply point to a particular bar to identify their own data value. The graph on the right of Figure 6 represents *grouped* data (reduced data), and specific data values are not obtainable for any individual student. Bright and Friel wanted students to determine "not only what information is presented in each representation but also whether identical information could be extracted from the different representations" (1998, p. 69). The researchers found that students had difficulty in making the translation from graphs of ungrouped to grouped data. As one student commented when looking at a graph for grouped data, "I don't know how you add this thing up" (p. 74). The central theme propounded by Bright and Friel is that "establishing connections or translations among representations is critical for developing understanding" (p. 82). While their analysis of results does discuss aspects of mode, median, and mean in the students' search for centers, Bright and Friel paid less attention to the ways in which variation reveals itself in the various graphical representations. The lack of attention to subjects' conceptions of variation is noteworthy, because many of the researchers' graphs, as well as questions about what is typical for the data set, seem to provide a natural context in which to look at variability in data presented visually.

Comment: Synthesis needed

In summary, research indicates that there are different levels of graphical understanding held by students, and that the type of graph makes a difference on what people comprehend. It may well be that the research on graphicacy represents a missed opportunity to look at variation in the context of visual data displays (Shaughnessy, 1997), but it can now be seen that a study about variation, if incorporating aspects of graphicacy, can benefit from consideration of what a person can comprehend from graphs.

Aspects of Data Handling: Averages

The concept of average is fundamental to statistical thinking. (Davis & Hersh, 1986). Watson and Moritz (2000c) point out that for a long time, the school curriculum focused almost exclusively on the arithmetic mean as being synonymous with finding an average. The notion of

average as representative of the data, including the use of measures such as mean, median, and mode, has only been emphasized in the American school curriculum in the past decade. Note that when talking about measures of central tendency, the term *representativeness* is used not as it was in the judgment heuristics discussed earlier. Kahneman and Tversky used the term as a label for a type erroneous intuitive thinking, while in the context of this section the word has a positive connotation for statistical thinking. Representativeness as it will be used in this context refers to a way of describing how well an average summarizes a set of data.

Work done in a study by Mokros and Russell (1995) of fourth, sixth, and eighth-graders showed that "students' notions of representativeness or typicality grow out of their everyday experiences and have a strong flavor of reasonableness and practicality" (p. 21). Students showed what they knew about average by constructing different data sets that could reflect a given average. Some of these construction problems were made richer through added constraints, such as the prohibition of using the actual average in the data set, or the mandatory use of preexisting data values.

For example, in the "Potato Chips" problem, the task was to put price tags on 9 bags of chips so that the typical, or average, price would turn out to be $ 1.38. An added constraint to extend the problem was to disallow any price tag from actually being $ 1.38. In the "Allowance Construction" problem, the aim was construct a distribution of allowances which, taken together, had an average of $1.50. An extension was to specify that 2 allowances had to be 75 cents, and 3 had to be one dollar. Thus, "the student's task was to create a large frequency distribution where some data was already placed" (Mokros, Russell, Weinberg, & Goldsmith, 1990, p. 4). Other problems asked students to interpret what was "typical" from graphs. In the "Allowance Interpretation" problem, a skewed, bimodal graph showing a number of students' allowances was presented. The task was to "use the data to determine the typical allowance as well as the highest amount that could be argued for" (Mokros & Russell, 1995, p. 24). A third type of problem involved understanding a weighted average. The "Elevator Problem"

essentially aimed at asking for the average weight of a group of ten people, comprising six men whose average weight was 180 pounds, and four women whose average weight was 125 pounds (Mokros et. al., 1990; Mokros & Russell, 1995).

Of the five predominant approaches to average identified by Mokros and Russell, two approaches - average as mode, and average as algorithm - were not associated with the notion of representativeness. Modal thinkers were easily able to construct data sets, since they saw average as the value occurring most frequently. However, "when they were not allowed to use the average value as part of their distributions, real difficulties were encountered" (Mokros et. al., 1990, p. 7). In general, students found it much harder to work from the average to the data than vice versa. The standard algorithm for finding an average doesn't go far in helping to solve construction problems, as is exemplified by one student whose first attempt at solving the "Potato Chip" problem was to take the desired average of $ 1.38, multiply that by 9, and then divide by 9.

> "When the interviewer asked if there's any way she could put some prices on the chip bags, she replied that she knew how to get an average, but had not yet learned how to find the 'numbers that go into an average'" (p.15).

The other three approaches - average as reasonable, or as midpoint, or as a point of balance - exemplified a sounder understanding of what it means for an average to represent a set of data. Average as "reasonable" was considered by Mokros and Russell to be significant for a more robust understanding about the concept.

When considering a data set, two fourth graders and two sixth graders demonstrated the key features of what the researchers mean about the notion of average as reasonable. First, the students relied on values that made sense in the *context* of the problem, in line with their own understanding of prices, allowances, or weights. Second, the students had some regard for the idea of average as roughly centered rather than precisely in the middle of the data, and that in some sense "high values must be countered by low values" (p. 9). Another important aspect of

the research was the finding of how strong an attraction symmetry had for some students, particularly those who gravitated towards a midpoint strategy (Mokros & Russell, 1995). The researchers concluded that premature introduction of the algorithm for finding the mean may in fact be impede a students' overall understanding of the ways in which the mean is or is not representative of the data.

To broaden childrens' narrow view of average, Mokros and Russell suggested it is important to "focus on describing and comparing data sets" (1995, p. 37). This suggestion was central to a study of eighty-eight students in grades 3 to 9 on the emergent ideas of statistical inference by Watson and Moritz (1999), as was the notion of average as a representative measure. The researchers asked students to compare the performances of the two classes who test scores were shown in frequency graphs (see Fig. 6). Students were asked which class did "better".

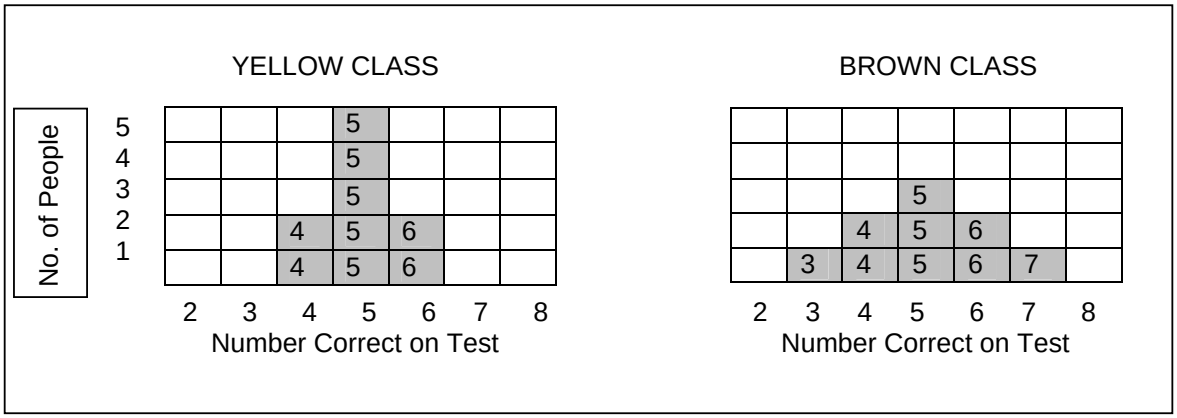

Fig. 6

The researchers classified student responses according to two strategies (numerical or visual) which were used either singly or together. For example, some students using a numerical strategy saw the graphs as a way to obtain the actual scores to calculate totals and means for comparing the two classes. Students using visual strategies commented on aspects like the symmetry or spread of the graphs. Some students "commented on both visual and numerical

strategies, but expressed conflict between them rather than viewing them as complementary related strategies" (Watson & Moritz, 1999, p. 155).

In the example shown above in Figure 6, the two classes were of equal size. Another similar example used classes of unequal size, with the result that proportional reasoning became increasingly important in analyzing the problem. The researchers concluded that students need more experience with a variety of data sets, and with tasks that allow for the representation of data graphically, and more experience with summarizing data with measures of central tendency.

Watson and Moritz (2000c) also explored students' longitudinal development of the understanding of average as mean, median, and mode. Their subjects were 94 students from grades 3 to 9, and the nature of their study allowed some initial participants to be involved over a four year period. Their results confirmed the trend suggested by Mokros and Russell (1995) that older students tend to "view data based problems in such a way that average is seen as a summary statistic which represents an entire data set" (Watson & Moritz, 2000c, p. 51). When students first learn the standard add-and-divide-by-total algorithm for the mean, they may come to associate this algorithm with the desired procedure whenever the concept of average is involved. That is, if a question asks "What's the average…?" then students may automatically think in terms of a mean, and attempt to apply the standard algorithm. Also, "when students do finally appreciate how to use the algorithm for the mean in a more difficult setting, they appear not to forget it" (p. 50).

When Mokros and Russell conducted their research on schoolchildren, they also gathered data on teachers from the same schools as those children. They asked the teachers the same questions about averages as the students, and found that, "in general, the teachers were more difficult to classify according to their predominant approach to problems than the students" (Russell, Mokros, Goldsmith, & Weinberg, 1990, p. 2). Teachers were also the subjects in study of the concept of average by Callingham (1997). She took a sample of 136

pre- and in-service teachers of students from kindergarten up through tenth grade, and administered a written survey. One question involved bar charts showing the results on the same spelling test for two groups of differing sizes (see Fig. 7)

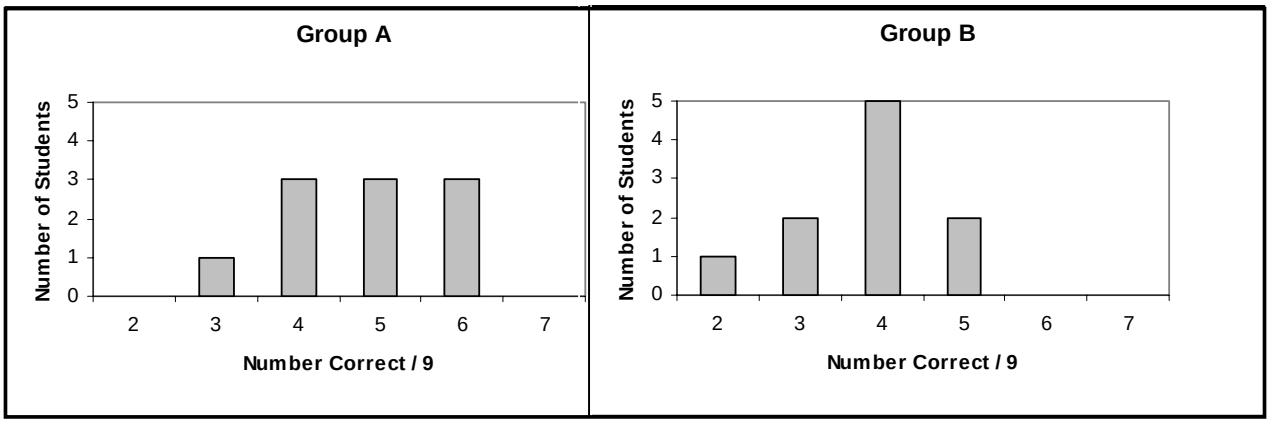

Fig. 7

The question asked in this case was to tell which group did better on the test, or whether the groups did equally well. With teachers, Callingham obtained similar results to those obtained by Watson and Moritz (1999) with students. The teachers used both numerical and visual strategies, and Callingham noted that these types of responses were much like those used by children in acquiring the concept of average. In discussing implications for teachers' professional development, she suggests that "when concepts of average are being developed, data should be presented not only numerically but also visually" (Callingham, 1997, p. 222).

This latter comment is important, because it highlights the conflation of the aspects of graphicacy and average as representative of the data. As we turn to look at some studies which synthesize different notions of statistical thinking, it becomes clear that a study on variation cannot properly be divorced from other key ideas such as average and graphicacy. In looking at Callingham's (1997) graphs above in Figure 7, or Watson and Moritz's (1999) graphs from Figure 6, questions about which class did better beg for variation to be noticed. However, those examples of research were more focused on tasks and questions which elicited conceptions of

the notion of average. Loosen, Lioen, and Lacante (1985) used pairs of sets of colored blocks of different heights, and the effect looked very much like a picture of Callingham's or Watson and Moritz's graphs. The question asked by Loosen et. al. of tertiary students was, "In which of the two sets do you judge the blocks as most unalike? In which set do the blocks have the greatest variation among themselves?" (1985, p. 3). The intention behind the task was to probe students' intuitive understanding of the heterogeneity in data, and to lay the groundwork for studying the standard deviation as a measure of how strongly the data strays from central tendency.

Comment: Synthesis needed

A person's facility with graphs, measures of center, and concept of distribution all may influence how one thinks about variation. In the next section, shades of these different but related notions can be seen, as attention is now focused on research which highlights the contexts in which variation was explicitly revealed or explored. The research on variation now unfolds in three broad contexts: data sets, sampling, and probability situations. Much of the research which follows exhibits a blend of statistical concepts, such as graphicacy, averages, or distributions. In a way, many of these studies draws on statistical skills, concepts, and intuitions related to those discussed earlier. The main difference is that the variation inherent in the situations is more apparent, and in many cases was the focus of the investigation itself.

Variation in Data Sets

Jones, Thornton, Langrall, Mooney, Perry, and Putt (2000) claim that "for students to exhibit statistical thinking, there is a need for them to understand data-handling concepts that are multi-faceted and develop over time" (p. 271). Data handling incorporates "organizing, describing, representing, and analysing data, with a heavy reliance on visual displays such as diagrams, graphs, charts, and plots" (Shaughnessy, Garfield, & Greer, 1996, p. 205). The kind of thinking implied by these statements includes attention not only to graphicacy and averages, but to spread, or variation, as well. Data handling implies not only a set of data to handle, but also a way to reduce the data while retaining the key features of the data set.

Friel et. al. (1997) mention the process of data reduction and the structure of graphs as factors influencing graph knowledge. They note that "data reduction is an essential part of analysis of the data; different graphs emphasize different degrees of data reduction" (p. 62). In assessing the components of data reduction, they used a problem that involved a stem-and-leaf plot (Fig. 8) of minutes taken by middle grade students to get to school.

Minutes to Travel to School

0	3 3 5 7 9
1	0 2 3 5 6 6 8 9
2	0 1 3 3 3 5 5 8 8
3	0 5
4	5

Fig. 8

The accompanying question was, "What is the typical time it takes for students to travel to school?" (Friel & Bright, 1996, p.6). The researchers concluded that "students were less likely to compute measures of center as part of their responses" (Friel et. al., 1997, p. 60). Evaluating the different responses, the researchers seem to validate the idea that attention to variation was an important part of an overall analysis of the question. They note that only a few students chose to use the mean or median in their responses. Instead, results showed that students responded in terms of clusters of typical times, or in terms of a range of numbers that occurred more often, or in terms of the mode. They rhetorically ask of the various methods,

> "Is one 'more appropriate' than another; do we want students to move beyond the use of the mode as a tool in this case to using clusters of data as a way of describing what's typical? If so, what is a 'good sized' cluster to be highlighting?" (p. 60).

Thus, in the context of the graphical aspects of data reduction, students need an awareness of the importance of both measures of central tendency *and* the spread of the data.

The research of Jones et. al. (2000) similarly allows for a natural relationship to emerge between conceptions of average and spread in graphical displays. In developing their framework for characterizing elementary children's statistical thinking, Jones et. al. used four

constructs that correspond well to those posited by Shaughnessy et. al. (1996): *describing, organizing, representing,* and *analyzing* (interpreting) data. The first construct, *describing,* invokes Curcio's (1987) idea of reading the data. *Organizing* involves reducing data using notions of center and spread. The construction and use of visual displays to show different organizations of a data set are at the heart of *representing*. *Analyzing*, relates to Curcio's other two categories of reading between and beyond the data. As an example of questions in the second construct, to assess children's thinking about *organizing* and reducing data using measures center and spread, the researchers showed children two graphs showing the number of friends who came for a visit over the course of a week (see Fig. 9).

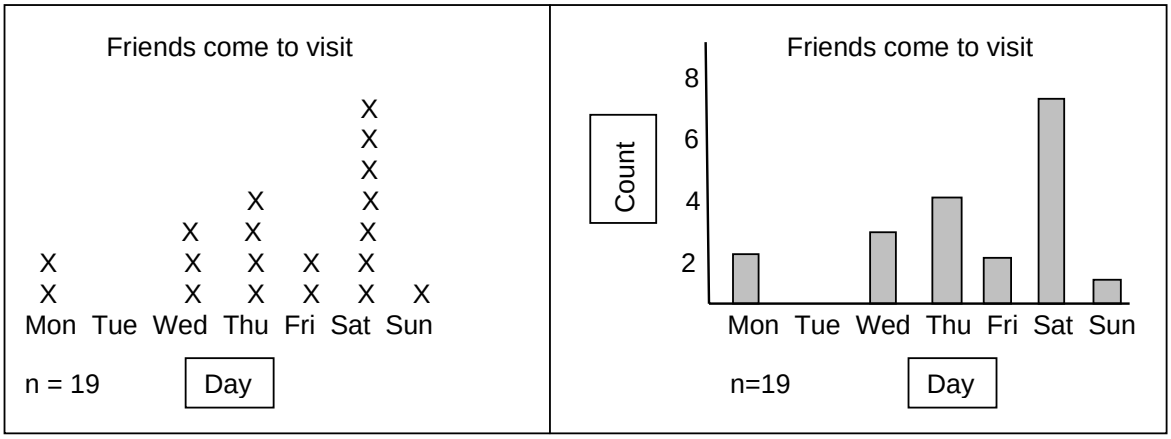

Fig. 9

The researchers used two variations of the same question to determine the children's understanding of center, "What's the typical number of friends who came to visit?" and "What's the average number of friends who came to visit?" (Jones et. al., 2000, p. 275). Also in this *organizing* construct, two sets of game scores were presented in tabular form, and a question dealing with variation was, "Which of these sets of scores have the greatest spread, or do they have the same spread?" (p. 275). To assess thinking about *analyzing* data, students were asked questions that involved inference or prediction, such as "About how many friends would you expect to visit during the next 4-week month?" (p. 277). Jones et. al. (2000) were guided by

the SOLO taxonomy, which is a hierarchical model of cognitive development. Student responses were therefore categorized by different levels of thinking according to the taxonomy. Level 2 thinking, for instance, incorporated some invented measures for dealing with center and spread. In relation to the question about the average number of friends who came to visit, one Level 2 student said, “Between 7 and 0. It’s somewhere there, but I don’t know” (Jones et. al., 2000, p. 295). By contrast, “students exhibiting Level 3 thinking consistently use quantitative reasoning as the basis for their statistical judgments and have begun to build valid conceptions of center and spread” (p. 298). One student at this level reasoned that although 19 friends visited during the given week, on other weeks the number of friends per week might vary from 10 to 15.

Reading and Pegg (1996) also used the SOLO taxonomy in their research on student responses to two open-ended tasks related to data reduction. Their subjects were one hundred and eighty students, thirty each from grades 7 to 12. Raw data on the length of 29 people’s feet were presented in list form for the first task, and graphical data on 28 test scores was given in the second task. The question for both tasks was similar. Students were asked to give the number or numbers that could be best used to represent the size of feet or score of students in that data set. The researchers identified a hierarchy of nine levels to categorize the responses. Overall, they noted, “students are far more likely to reduce the data using measures of central tendency than [measures of] dispersion irrespective of the data presentation” (Reading & Pegg, 1996, p. 194). They also found that some of the higher-level responses indicated “the use of both measures of central tendency and dispersion and the use of the features of the data in an attempt to establish a link between the two” (p. 190).

Another approach in looking at students’ statistical thinking focused on the concept of distribution. Mellissinos, Ford, and McLeod (1997) noted that although previous studies identified some ideas on how students reasoned about the concept of average, “they have revealed little about how students make sense of an average in the context of the distribution

that it represents or summarizes" (p. 176). Using the "Potato Chip" problem of Mokros and Russell (1995) described earlier, Mellissinos et. al. interviewed a middle-school student. Their results supported the findings by Mokros and Russell (1995), in which the subject did not understand what it meant for an average to be representative of a set of data. What distinguishes the work of Mellissinos et. al. (1997) is their focus on an understanding of distribution. They write, "One reason for the student difficulties may be that students have not learned to think about the mean as a representative measure of a distribution" (p. 179). Mokros and Russell's notion of average as a representative measure involves capturing a range and distribution of a set of data, but Mellissinos et. al. caution that an understanding of the distinction between the terms "distribution" and "data set" is necessary to establish and interpret representativeness. According to Mellissinos et. al. (1997),

> "A data set is a collection of measurements of one or more characteristics (of objects or people). A distribution is an attribute of a data set that communicates how measurements is a data set are distributed across its range of values" (p. 179).

Mellissinos et. al. state conclude that "without a clear idea about distribution, it is difficult to make inferences about student notions of representativeness" (1997, p. 179).

Mellissinos also researched how the notions of a distribution's shape, center, and spread interact (Mellissinos-Lernhardt, 1999). This time she used the "Potato Chip" problem with college students (not necessarily preservice teachers), and another task that involved the pulse rates of a set of 30 people in the same age group. One task used both a table and a histogram to present the 30 pulse rates. After first being told that the typical pulse rate for the given age group was around 65 to 69 beats per minute, students were asked if they thought that the pulse rates for a given set of people were typical. Results showed that, for some students, the ability to interpret a mean did require some concept of the range of possible values. In the Potato Chip task, one student expected variability in the prices for bags of chips, but did not "have a sense of how much variability would be possible for the situation" (Mellissinos-Lernhardt, 1999,

p. 6). In the Pulse Rate task, the same student

> "relies too heavily on the mean to decide whether the group is typical. She does not take into consideration how spread out the pulse rates are, how the data cluster and whether there are any outliers" (p. 7).

It thus did seem that some students had some awareness that only looking at the center does not capture the whole picture. Mellissinos reiterates that the while many educators promote the mean as representative of a distribution "the concept of distribution relies heavily on the notion of variability, or spread" (p. 1). Certainly the characteristics of a distribution can rely on representing data through summary statistics, but graphs also provide a representation. Hence, Mellissinos' research highlights the connections in statistical thinking between conceptions of graphicacy, centers, and variation, through the common unifying theme of distribution.

Shaughnessy (1997) discusses the work of middle school students on data sets stemming from the weather pages of a local newspaper. Students were gathered in groups of four to five, and each student was given a days' weather page so that their small group had a small set of consecutive days' worth of data on the weather. He observed and noted the types of questions that students came up with, and the approaches they took in analyzing their own questions. In their analysis students became aware of the kinds of variation inherent in weather data. For example, Shaughnessy wrote that "quite a discussion ensued as to why there was such variability for the coldest time of day" (1997, p. 12). The theme of weather was also used by Torok and Watson (2000), who explored concepts of variation with sixteen students, four each from grades 4, 6, 8, and 10. The two researchers framed their weather task using the concept of average. Students were told that the average maximum daily temperature for Hobart, Australia, over the year was 17 degrees Celsius. Questions included whether or not students thought all the days of the year had that maximum temperature of 17 degrees, and what the maximum temperature might be for 6 different days of the year. They also asked students for likely ranges of temperatures, such as what the highest and lowest maximum daily temperature might be for the month of January, July, or over the course of the year. Torok and

Watson (2000) found that their “suggested developmental levels of understanding variation are related to increasing grade level” (p. 163).

Shaughnessy and Pfannkuch (2002) have found that the data sets for the Old Faithful geyser provide an excellent context for highlighting the role of variation in statistical analysis. They describe a classroom exploration in which students were first given one day’s worth of data for the number of minutes between successive eruptions of the geyser. The question which threaded through this investigation was about how long one should expect to wait between eruptions of Old Faithful. Students were then told to represent this data in a graph of their choice. Then they were given several more days’ worth of data and asked to graph it as well. Some of the students used boxplots, which do demonstrate the range of the data but which mask the nature of variation which can be seen when the data is plotted over time. Other students used histograms, which reveal the underlying distribution. At first, many students just make an initial prediction based on measures of central tendency (such as the mean or median), which also disregard the variability in the distribution. Shaughnessy and Pfannkuch (2002) point out that

> “Students who attend to the variability in the data are much more likely to predict a range of outcomes or an interval for the wait time for Old Faithful, such as ‘Most of the time you’ll wait between 50 to 90 minutes,’ rather than a single value of 70 minutes” (p. 5).

Also, when students first look at one day’s data, and then look at several days’ data, they may see the variation across days as well as the variation within a day. This extends Curcio’s (1987) analysis to what these researchers call looking “behind the data” (p. 6).

The point of the research on this section is that questions about data sets can indeed be shaped to explore student understanding of variation. The trend in the research discussed so far is to recognize the importance of blending of variation with several other statistical concepts such as graphicacy, distributions, and averages. In addition to the contexts of data sets, students’ conceptions of variation can also be studied within the context of sampling, which is

Comment: Synthesis needed

next discussed.

Variation in Sampling Situations

As mentioned in the previous chapter, the term *variation* and its different linguistic forms come with many related meanings. In sampling situations, variation appears in the differences among repeated samples drawn from the same population. Samples also vary in the degree to which they represent their parent population. For example, is the population is all students at a high school, using the same opinion poll on two different groups of students is likely to produce to different poll results. Thus, sampling situations can invoke many levels of meaning when we consider variation.

Within the context of sampling, "as sample size increases, the statistics of a sample become less variable and more closely estimate the corresponding parameters of the population from which the sample was selected" (Well, Pollatsek, & Boyce, 1990, p. 289). This is due to the Law of Large Numbers, yet Kahneman and Tversky (1972) found, through tasks like the Hospital Problem mentioned earlier, that many people are unaware of how sample size influences variability. Well et. al. (1990) point out the objections of other researchers, who criticized some of Kahneman and Tversky's problems as being too difficult for subjects to fully comprehend.

Some research has shown that people can be correctly influenced by sample size. For example, in a question that asked which of two samples of different sizes would better estimate some characteristic of a population, Bar-Hillel (1982) found that over 80% of subjects correctly chose the bigger sample. To gain further insight into how well people understood the effects of sample size on the variability of the mean of the sampling distribution, Well et. al. (1990) posed a series of questions to undergraduates,

> "in some cases asking subjects to judge which of two samples was more likely to fall closer to the population mean and in others, asking them to judge which of two samples was likely to deviate more from the population mean" (p. 292).

One context used by Well et. al. was the average height of American males. Students were told that the national average height of 18-year-old males is 5 feet, 9 inches. They were also told that at Post Office A, 25 men registered for the draft each day for a year, while the number of men registering at Post Office B was 100 men per day for a year. At each Post Office, the average heights of men per day was computed. As an example of a question pertaining to the tails of the distribution, students were asked which Post Office would have recorded more days on which the average height was 6 feet or more. The corresponding question pertaining to center of the distribution was identical, except that "6 feet or more" was replaced by "between 5 feet 6 inches and 6 feet" (Well et. al., 1990, p. 297). Results indicated that people used information on sample size more accurately when dealing with centers of distributions rather than the tails. However, even when the subjects had received instruction on sampling distribution, "many of them still did not understand how sample size influenced the variability of the sample mean" (p. 310). This research does highlight the importance of understanding a distribution, a point also brought out in the research of Mellissinos (1999).

Some of the questions used by Well et. al. are similar in spirit to the Hospital Problem and to a question adopted by Watson, Collis, and Moritz (1995). These latter researchers interviewed a subset of twelve students from 171 girl subjects from grades 3, 5, 7, and 9. They intended to explore the general notion that small samples are more likely to have extreme results than large samples, with this question:

> "The researchers took a random sample from each school: 50 children from the city school, 20 children from the country school. One of these samples was unusual: it had more than 80% boys" (p. 6).

Students were then asked whether the unusual sample was more likely to have come from the city or country school. Not one of the twelve students was able to give a suitable justification for their response. The same question was also used by Watson and Moritz (2000b), who asked 41 students from grades 3, 6, and 9. They found that only six students could give an adequate explanation. Many of the other responses made it seem as though the choice of large or small

sample was almost a random decision, “with the reasons given for choosing the small samples also being given for choosing the large one” (Watson, 2000a, p. 122). The original Hospital Problem can be theoretically modeled by a binomial distribution, while this modified version above involves a hypergeometric distribution (since children are chosen without replacement). In either case, many students do not recognize that “the smaller sample is more likely to give an extreme or biased result” (Watson & Moritz, 2000b, p. 66).

Fischbein and Schnarch (1997) included the original Hospital Problem in their study on the evolution of probabilistic, intuitively based misconceptions. Of eighty students in grades 5, 7, 9, and 11, plus 18 prospective teachers, only one ninth grader suggested that the smaller hospital would have the more extreme result. In addition, the percentage of respondents who thought the results would be about the same for both hospitals actually increased with age, from 10% of the fifth graders all the way up to 89% of the prospective teachers. Related to the heuristic of representativeness mentioned earlier, the basic misconception here is that sample size doesn’t affect variation. The researchers note that “this misconception developed with age in a surprisingly regular manner” (Fischbein & Schnarch, 1997, p. 101), meaning that more people committed the misjudgment as the age of the subjects increased.

Watson (2000a) gave 33 preservice secondary mathematics teachers the Hospital Problem, and categorized her results by whether the respondents based their solutions solely on intuition, on a mathematical argument, or on a mixture of approaches. She found the success rate of 55%, while not surprising in comparison to results of Kahneman and Tversky’s (1972) study of tertiary students, was not related to her subjects’ prior formal mathematics experience. Moreover, she notes that “it is disappointing that so few naturally mixed intuition with an attempted mathematical justification in solving the hospital problem” (Watson, 2000a, p. 134).

The Hospital Problem and similar questions not only illustrate the kinds of intuitions and use of heuristics shown by Kahneman and Tversky (1971, 1972), but also allow researchers to

investigate students' understanding of the relationship of a sample to its underlying population. As Watson and Moritz put it, such questions allow researchers to "investigate students' understanding of the effect of increased sample size in increasing the reliability with which the sample represents the populations" (2000b, p. 50). Watson and Moritz (2000a) wished to determine if "students recognize the tension between efficiency of small sample sizes, and the reliability of larger sample sizes" (p. 6). They asked 2040 students from grades 3 through 11 whether someone should choose to buy one car in favor of another based on the opinions of a few friends, on the results from a statistical report on 400 cars of each type, or whether both sources of information were equally valid. More students chose the response that both sources were equally valid over any other response, although the older students more successful identified the greater reliability inherent in the larger sample. The researchers (Watson & Moritz, 2000a) noted that,

> "Some students believe that any sample however small is representative while others believe that 'larger' is always 'better' to achieve representativeness, without regard to increasing difficulty and cost of data collection" (p. 31).

This tension is further revealed in some of the subsequent research on sampling which was undertaken in the context of conducting surveys, or polls.

Jacobs (1997) conducted a study using fourth- and fifth-grade subjects, in which childrens' informal understandings about sampling issues were investigated. Specifically, Jacobs was concerned about students' perceptions of sampling methods, and in students' distinctions between those methods that led to representative samples versus those that led to unrepresentative samples. Some of Jacobs' questions were in the context of taking a survey to predict how many schoolchildren will purchase a ticket for the school raffle. Other questions were in the context of taking a survey to determine how many of the city's schools were recycling. In each context, subjects were presented with a variety of sampling methods to evaluate in conducting the survey, such as simple and stratified random sampling methods, self-

selected methods, and restricted methods (which used select groups of the population who would be more likely to skew the results in a certain direction).

Jacobs found that children evaluated these sampling methods by focusing on the potential for bias, fairness, practical issues, or the results produced by the method. For example, “some children…were able to identify potential bias with restricted and self-selected sampling methods and to recognize a lack of potential bias with random sampling methods” (Jacobs, 1997, p. 12). In focusing on fairness, however, some students were not impressed by the simple random methods which ensured that everyone had the same chance of participating in the survey. They were “not thinking of fair in the probabilistic sense but rather in the affective sense of how the participants (or non-participants) felt about having the opportunity to participate in the survey” (p.12). This interpretation of fairness led subjects to want members from all types of subgroups in the sample, which meant that stratified random sampling was a favored method for these subjects. Some of Jacobs’ subjects also focused on the possibility of extreme outcomes, even though the chance of those outcomes occurring was low. Other researchers have claimed that the focus on extreme outcome is suggestive of the outcome approach (Shaughnessy, Watson, Moritz, & Reading, 1999), because even though an outcome might have a small probability, the outcome still *could* occur. Lastly, students assumed a correspondence between the results produced by the sampling method and the results expected prior to sampling. That is, “if the results corresponded with what was expected, then it was an appropriate sampling method because it got the ‘right results’ “ (Jacobs, 1997, p. 14). The converse also held, so that sampling methods which produced different results from what was expected are judged to be incorrect methods.

Statistical inference involves using the data at hand to make predictions about the population from which those data were taken. Jacobs rightly notes that “statistical inference is almost by definition imperfect – all sampling introduces some error” (1997, p. 2). The context of media polls has been used to research how students expect, notice, and understand the

variation inherent in sampling (Watson, Collis, & Moritz, 1995; Watson, 1997; Watson, 1998; Watson & Moritz, 2000a; Watson & Moritz, 2000b). Polls invite people to consider the effect of sample size on variation, as well as the variation naturally arising from polling different sample of the same size. One task, involving polls on handgun use, generalized from a sample of 2508 Chicago high school students to make a claim about all high school students in America. Another poll involved listeners who phoned a youth radio station to voice opinions on drug use. The research tasks asked whether or not the samples for these polls offered a reliable way of finding out public opinion throughout the country. Watson and Moritz (2000b) used the SOLO taxonomy to analyze and interpret results. They hypothesized a model of student development of concepts of sampling which suggests that

> "As students begin to acknowledge variation in the population, they recognize the importance of sample selection, at first attempting to ensure representation by predetermined selection but subsequently by realizing that adequate sample size coupled with random or stratified selection is a valid method to obtain samples representing the whole population" (p. 63).

The concepts of variation and representation are intrinsic to the task of making inferences about populations from sample data, and the tension between these two concepts "always exists in a sampling situation" (Shaughnessy et. al., 1999, p. 7).

Rubin, Bruce, and Tenney (1991) agree that a key to mastering statistical inference is to balance sample representativeness (the way in which a sample often has characteristics that are identical to the parent population) with sample variability (the idea that different samples from a single population are often not identical). To investigate these concepts, Rubin et. al. (1991) interviewed a dozen high school seniors who had never taken any statistics courses. The researchers used a question in which the population was known, and repeated samples could be drawn. In the Gummy Bears problems, students were told that packets of candy were filled with 6 Gummy Bears per packet. These candies were packaged after being drawn from a large vat containing two million green and one million red candies. Students were first asked

about the number of green candies they thought would be in their own packet; then they estimated how many packets out of 100 would have that same number of green candies. "Finally, we asked them to specify the entire distribution by answering the questions, 'How many kids out of 100 had N green gummy bears in their packet?' for N = 0 through 6" (Rubin et. al., 1991, p. 5). This is an excellent question to get at variability in repeated samples of a fixed size. For the first question, all twelve subjects said that they would expect 4 out of the 6 candies in their own packet to be green, and their explanations indicated that they were using a ratio approach to this question. However, "when asked if every kid's packet would contain 4 green Gummy Bears, all of the students knew that there would be variation among samples" (p. 5). Some students felt a need to determine a cause for this variation, suggesting that the candies might have gotten stuck together; others clearly felt that any number other than four, while possible, was an example of a flaw in the sample. In looking at the distribution of 100 packets, students consistently overestimated the frequencies near the middle of the distribution, and underestimated the frequencies near the tails. "No student's distribution contained a peak at a point other than 4G, 2R, and…only two students allowed the possibility of a category being empty" (p. 7), the researchers noticed. For this task, students seemed overwhelmingly influenced by the notion of sample representativeness.

Rubin et. al. (1991) took the same subjects and asked them another question that involved the underlying binomial structure in the Gummy Bears problem. The subjects were asked to imagine that there were only enough lockers at a school for half the children. In order to determine who would or would not receive a locker, the school principal put slips of paper were put into a bowl. Half of the slips permitted the holder to a locker, and half denied the holder a locker. On the first day of the drawing, three friends pulled three slips and they all got lockers, but on the next day three more friends pulled three slips and all were denied a locker. Rubin et. al.'s subjects were then asked what kind of evidence should be gathered to determine whether or not the slips were properly mixed in the bowl.

Students' responses for estimates of needed sample size ranged from 50 to 500. For example, one student suggested that if the evidence showed 90 Yes Lockers and 5 No Lockers the first day, combined with 5 Yes Lockers and 90 No Lockers the second day, that would be good evidence of unfair mixing. The researchers comment that "students consistently chose samples that were extremely unlikely, i.e. likely to occur much less often that 1 out of 1000 times" (Rubin et. al., 1991, p. 9). Students were reasoning in this case as though "sample variability were the most relevant fact about sampling" (p. 11). These students insisted on a very convincing sample before inferring anything about the population. Thus, in two different contexts, Rubin et. al. were able to powerfully illustrate the twin ends of the continuum between sample representativeness and sample variability. On one end, sample representativeness is the idea that a sample may have characteristics that are identical to the parent population. On the other end, sample variability is the idea that different samples from the same population are not all identical and therefore do not exactly match the population. They conclude by noting that students "lack experience thinking in terms of a *distribution of samples* generated from a particular population" (1991, p. 12, italics added).

Shaughnessy (1997) shares anecdotes about a task in which repeated samples of M&M candies, each of size 20, were drawn from a population known to contain 40% brown candies. He notes that "no one has yet said 'you will get 8 browns every time" (p. 7). His point is that the idea of a range of likely values is accessible to students. Moreover, questions about the likely spread of values in a data set, or about the likelihood of a certain spread reoccurring by repeating the experiment, are good ways to get at the variability inherent in a resampling situation (Shaughnessy et. al., 1999; Shaughnessy, 1997).

The Gumball Task on the 1996 NAEP was a missed opportunity to look at student responses to questions about variation (Shaughnessy et. al., 1999). In the NAEP Gumball Task, students were shown a picture of a gumball machine and informed that the mix of 100 gumballs inside comprised 20 Yellow, 30 Blue, and 50 Red. The question asked students to

predict the number of red gumballs that would occur in a sample of size 10. The percentage of student responses that fell in the top level of the scoring rubric for this question was quite low (Zawojewski & Shaughnessy, 2000). A troubling aspect of this task is that the question "tries to tap children's understanding of centers but does so in a context which more naturally deals with spreads" (Shaughnessy et. al., 1999, p. 8). Notice for instance, that while the NAEP question is much akin to the first question asked by Rubin et. al. (1991) in the Gummy Bear problem related earlier, Rubin et. al. extended the line of questioning to get at the variability of results of repeated samples.

Other researchers explored ways of expanding the Gumball Task so that it offered respondents a chance to demonstrate what they knew about variation (Torok & Watson, 2000; Reading & Shaughnessy, 2000; Shaughnessy et. al., 1999; Shaughnessy & Ciancetta, 2000). In the amendments to the original 1996 NAEP gumball item (later called the Candy Task for research in America and the Lolly Task in Australia) several different ways of framing the task were created. The situation was changed to a repeated sampling problem in which five samples, or pulls, of size ten were drawn (with replacement). The RANGE version of the question, asked for the lowest and highest number of reds that would result from the five pulls (Shaughnessy et. al., 1999). The CHOICE version provided five preselected lists of possibilities for consideration, and the LIST version allowed the respondents to simply write down the five estimates for the number of reds in each pull. In all cases, respondents were asked to provide explanations for their responses. The 324 subjects, from grades 4, 5, 6, 9, and 12, were randomly assigned to one of the three versions of the task. Results were categorized on the basis of how the students' answers reflected their sense of center as well as their sense of spread. Students' sense of center was classified either LOW – FIVE – HIGH, while a NARROW – REASONABLE – WIDE scheme was used in classifying responses according to spread. Range widths from 2 to 6 were considered reasonable, with anything below being narrow and anything above being wide.

As an example of a response classified as “FIVE, NARROW”, one 12th-grader responded, “In this circumstance 10 lollies were pulled out, 50% of 10 = 5 therefore 5 red lollies would usually be pulled out” (Shaughnessy et. al., 1999, p. 18). The researchers note that “the students who give a NARROW response to the candy problem are not understanding the nature of variability in this problem” (p. 14). The students who were more successful usually attended to the population proportion of reds rather then merely to the total number of reds in the population. Other students seemed to use the outcome approach, and thus had a WIDE range because they felt anything was possible. For example, another 12th-grader’s response (classified as “FIVE, WIDE”) reads, “There is always the probability of getting no red lollies, on the other hand, there is a possibility of getting all the lollies red” (p. 18). This student focuses on the extreme outcomes, believing that anything *could* happen. A subset of the younger students in the study were given the Candy Task both before and after they did the actual experiment in the classroom. Students took turns making the various pulls, recording the results, and then remixing the contents of the container. The researchers found that there was “considerable improvement in the students’ responses after they actually did the experiment” (p. 15).

Reading and Shaughnessy (2000) extended the Candy Task, altering either the number of pulls or the sample size. For example, they asked students for the numbers of reds if six people each pulled out samples of size 50 with replacement. This allowed for an exploration of responses for an increased sample size. They also asked students to describe the results of 40 pulls, each of sample size 10, and then asked students to graph the results for 40 pulls. Finally, they altered the candy mixture itself from 50R, 20Y, 30B to 70R, 20Y, and 10B. Six elementary and six secondary students were interviewed on these tasks. In all cases the students were asked to provide an explanation for their responses. Results showed that students were better at describing reasons for their responses when talking about centers than when talking about variation. Also the researchers found that “the LIST form of the question appears to give more information about variability than the CHOICE or RANGE versions,” and

that “it may be hard for students to describe variation with only six handfuls” (Reading & Shaughnessy, 2000, p. 7). Another finding was that the different mix of colors did not seem to affect student ability in predicting outcomes, although “it appears to be more difficult for students to justify their responses with the 70% mix” (p. 8). The researchers revised their protocol to ask students to imagine pulling 100 samples of size 10, and to draw a histogram for the frequency of reds in each sample. Reading and Shaughnessy suggest that a computer simulation would be useful to display to students. Researchers could then investigate whether students would want to revise their own suggested histogram after seeing simulation results.

Comment: Mike and Chris use the word “histogram” on Page 7.

In an exploratory study involving four students each from grades 4, 6, 8, and 10, Torok and Watson (2000) used the same expanded form of the Candy Task that Reading and Shaughnessy (2000) used. Students were asked if the results of their pulls were surprising, and were given an opportunity to modify their earlier answers after doing the experiment. The strongest factors which differentiated students’ responses were “the extent to which variation was acknowledged and…the recognition and use of the proportion concept to describe individual outcomes” (Torok & Watson, 2000, p. 153). These factors gave rise to four hierarchical levels that comprised a model for categorizing student reasoning. At the lowest level, subjects acknowledged variation but focused on individual outcomes. The four students at this level were easily swayed by experimental results. At the other end of the hierarchy, the two students in the highest level showed a high level of proportional thinking, balanced by a “very good and consistent appreciation of variation” (p. 160). These students were only moderately influenced by experimental results.

It seems fair to wonder what role graphicacy plays in the Candy Task. Torok and Watson (2000) expressed surprise at the general lack of facility of students in generating graphs to show the outcomes of 40 draws of 10 candies each. Shaughnessy and Ciancetta (2001) asked 31 secondary mathematics students to graph the expected outcomes of 100 draws of 10 candies each. Shaughnessy and Ciancetta note that “in general, constructing a

bar graph to represent the results proved a difficult task for these students" (2001, p. 13). Torok and Watson (2000) propose an amendment to the Candy Task protocol which would ask students to fill in a partially completed bar graph, suggesting that "this would be likely to reveal more about students' conceptions about the clustering of results around their expected values" (p. 164).

The Candy Task seems well-suited for investigating not only the effects of sample size, but also on the effects of increasing the number of samples of a fixed size. Moreover, the kinds of questions which have been asked require facility in reasoning about centers, spreads, and also a sense of graphicacy. There are also ways in which the subjects' sense of distribution can be tapped. Saldanha and Thompson (2001) had groups of students draw random samples from populations whose composition was not revealed to the students. After noting that the variability among samples made it difficult to make claims about the population composition, the students wanted to look at collections of samples. Thus, "each group drew 10 samples of equal size from a population of objects and the class investigated how these collections, as a whole, were distributed" (Saldanha & Thompson, 2001, p. 2).

By incorporating so many different aspects of statistical thinking, the sampling items presented thus far seem very versatile as a way to investigate students' thinking about variation. Research shows that sampling environments provide opportunities to look at the effect of sample size on variation, as well as the way that samples of the same size differ from one another and provide different pictures of the underlying population. A last context for looking at conceptions of variation is in probability situations, and this context is described in the next section.

Comment: Need summary

Variation in Probability Situations

Much of the previous research on sampling is colored by probabilistic thinking. One can imagine being asked for the probability of getting a certain number of red candies in a sample, or the chance of being selected to participate in a survey. For example, I chose to separate

Truran's (1994) research on children's understanding of variance from the body of literature discussed so far in this paper, because his study was framed in terms of "one probabilistic situation" (p. 2). Truran used colored balls in an urn, much akin to the Candy Task, but he only had two greens and one blue in the urn. His subjects, four girls and four boys in each of grades 4, 6, 8, and 10, drew one ball at a time with replacement. Truran based a series of questions on this format, "If we did this again m times would you be surprised if you got n greens / blues?" (p. 3). This is similar to the Candy Task, except that here the sample size is one. This distinction makes the probability aspects of this task more transparent than the sampling aspects. Still, in asking first about 9 draws and then about 50 draws, Truran's aim was to find out about students' conceptions of variation, and in particular to see what range of results the students would consider normal. It is interesting to note that while the protocol asked for students to provide a specific number they would find surprising, in fact some students explicitly talked about *ranges* of surprising values without being prompted. Truran notes that almost all of the subjects "had some awareness that extreme numbers would be surprising," and seven of them "distinguished 'surprising' from 'very surprising'" (1994, p. 7). He also noted some reliance on the availability heuristic, and claimed that the students' naïve understanding of variance depended on their number sense and their facility in computation.

Shaughnessy (1997) mentions that when students are given a probability question that involves the likelihood of a single event, some students may try to superimpose the idea of a sample on the problem when none exists. For instance, a question was posed to middle school teachers in which a fair coin was flipped five times. Teachers were asked what is more likely to occur, A) HTTHT, or B) HHHHH. Some of the teachers responded that "'In a small sample anything is possible' and "the long term results gravitate towards A),' as if there were an ongoing sample" (Shaughnessy, 1997, p. 3). In fact, there is no sample at all in this question. There is a sample space in the probabilistic sense, but no actual sample from a population is being drawn (Shaughnessy et. al., 1999). Questions like these lead subjects to "focus on single

outcomes, rather than a range of possible outcomes" (Shaughnessy, 1997, p. 3). Konold (1995) used simulations to look at many trials of five flips, where the number of trials corresponded with what he called the sample size. He found that students could see the kind of variability among repeated samples of the same size. Thus, the outcomes in one sample of size 1000 (that is, completing 1000 trials, with each trial constituting five flips) will vary from those in another sample of size 1000. Konold claimed that "the different outcomes of each repetition reveal the variability inherent in the sampling process and give some sense of the magnitude of that variability for the given sample size" (1995, p. 209).

Comment: Omitting the Konold "Confessions of a Coin Flipper" and also my connection to Cereal Box spinners.

Shaughnessy and Ciancetta (2001, 2002) used a pair of spinners with 31 secondary mathematics students. Each spinner was half black and half white. First the students were asked a pure probability question: "A player wins a prize only when *both* arrows land on black after each spinner has been spun once. Jeff thinks he has a 50-50 chance of winning. Do you agree?" (Shaughnessy & Ciancetta, 2002, p. 2). Then, students were asked to predict the number of times out of ten spins that both spinners would result in black. After predictions were made, students gathered data in sets of ten spins each. After each successive set of ten spins, students were given an opportunity to revise their predictions. This same protocol was also repeated with a pair of spinners in which the first spinner again was half black and half white, while the second spinner was one quarter black and three quarters white. For some students, there was conflict "in trying to resolve what their 'theories' would predict, and the variability in their sample data" (Shaughnessy & Ciancetta, 2001, p. 12). Prior to gathering data, very few subjects actually listed a sample space for these problems. However, actually performing the experiment, gathering the data, and "seeing the variation in repeated samples of ten trials, led a number of our students to construct the sample space for the spinner problem" (Shaughnessy & Ciancetta, 2002, p. 5). These interview results support the connection between the two concepts of the sample space in probability and the expected variation in values of a random variable. They note that

> "The conceptual root of the pedagogical power that we gain from having students conduct simulations is the connection that they can make between the observed variation in data in repeated trials of an experiment, and the outcomes that they expect based on a knowledge of the underlying sample space or probability distribution" (p. 6).

Thus, probability experiments do seem to offer promise as a viable context for gathering data on people's conceptions of variability. Repeated trials of a probability experiment can focus attention on the variation inherent in the outcomes, rather than just on the expected value for any particular outcome.

Comment: Summary needed

Conclusion

Statistical thinking is influenced by a host of different factors. Some research deemed most salient to a study of conceptions of variation has been discussed in this chapter. This includes research on randomness, intuitive stochastical thinking, graphicacy, and averages. A lack of appreciation for randomness appears to translate into an inadequate view of variation, and to diminish the promise of statistical inference. Naïve and intuitive views of uncertainty can result in misjudging probabilistic and stochastical situations, including those situations where reasoning about variation is crucial. Graphs are a powerful tool for reducing data, and different graphs convey different degrees of information about center and spread. Also important for data handling are measures of central tendency. All of these ideas relate to conceptions of variation. For example, if students have difficulty reading graphs, then they will likely be unable to notice variation as it is revealed graphically. If they are unable to find a representative measure for a data set, then they will be unable to talk about variation as a spread around the center of a distribution.

In recent years it has become clear that studies crafted specifically to elicit responses on variation can be embedded in contexts which include (although certainly not limited to) data sets, sampling, and probability situations. Although some research has been conducted on students' conceptions of variation in these three contexts, one of the larger gaps in the literature concerns research specifically looking at what teachers know about variation. Shaughnessy

(2002) writes that “we are not aware of any research studies that have dealt specifically with teachers’ conceptions of variability” (p. 2), and this paucity of research also holds true for preservice teachers.

This study proposes to research preservice teachers’ conception of variation, and it aims to do so in the three contexts of data sets, sampling, and probability situations. The next chapter profiles the philosophical orientation and conceptual framework guiding the study.