13. STUDENTS ANALYZING DATA: RESEARCH OF CRITICAL BARRIERS

Clifford Konold, Alexander Pollatsek, Arnold Well
University of Massachusetts, Amherst
Allen Gagnon
Holyoke High School

INTRODUCTION

In describing the work of the nineteenth-century statistician Quetelet, Porter (1986) suggested that his major contribution was in persuading some illustrious successors of the advantage that could be gained in certain cases by turning attention away from the concrete causes of individual phenomena and concentrating instead on the statistical information presented by the larger whole.

This observation describes the essence of a statistical perspective−−attending to features of aggregates as opposed to features of individuals. In attending to where a collection of values is centered and how those values are distributed, statistics deals with features belonging not to any of the individual elements, but to the aggregate that they comprise. Although statistical assertions such as “50% of marriages in the U.S. result in divorce” or “the life expectancy of women born in the U.S. is 78.3 years” might be used to make individual forecasts, they are more typically interpreted as group tendencies or propensities. In this article, we raise the possibility that some of the difficulty people have in formulating and interpreting statistical arguments results from their not having adopted such a perspective, and that they make sense of statistics by interpreting them using more familiar, but inappropriate, comparison schemes.

Propensity

It was not until the nineteenth century that statistics surfaced as a discipline and the idea took root that certain variable data could be described in terms of stable group tendencies, or, as Quetelet (1842) referred to them, “statistical laws.” In this article, we refer to these group tendencies as “propensities.” This should not be confused with its meaning in Popper’s propensity theory of probability, in which a relative frequency is considered a measure of a stable, physical property of a chance set-up (Popper, 1982). By group propensity, we mean the intensity or rate of occurrence (an intensive quantity; Kaput & West, 1993) of some characteristic within a group composed of elements that vary on that characteristic.

For example, to say that 75% of teenagers at a certain school have curfews indicates the tendency of students in that *group* to have a curfew. The group comprises in this case both students with and without curfews. Saying instead that 350 students at that school have a curfew creates a group within which there is no longer any variability: that is, all 350 students in the group have a curfew. Given that we do not know the size of the school, the 350 says almost nothing about the tendency of students at the school to have a curfew--in this sense it is not a propensity. Thus, not all properties of aggregates are propensities.

Nonstatistical comparisons

Comparing the size of two sets and comparing two individuals with respect to some attribute are the types of comparisons that will be described here as "nonstatistical." For example, we could measure the height of two individuals and then declare the one with the larger measurement to be the taller of the two. Here we are measuring or classifying each individual on some attribute and then comparing the two with respect to that measurement. Because there is no variability involved (assuming each individual was measured once) we consider this a nonstatistical comparison. Elaborating on the example provided above, we could count the number of males and females in a particular school who have curfews and then claim that more females than males have curfews. In this case, we first assign individuals to groups based on their classification regarding some attribute(s), count the number in each group, and then compare the sizes of those groups. Again, we do not consider this to be a statistical comparison because members of each group are of the same type with respect to the quality of interest; thus, there is no variability within groups.

This paper reports the results of interviews with high-school students who had just completed a year-long course in probability and statistics. Using analysis software they had learned as part of the course, these students had difficulties formulating valid statistical comparisons. We argue that their failure to make appropriate use of statistical measures such as medians, means, and proportions is due in part to their tendency to think about properties of individual cases, or about homogeneous groupings of such cases, rather than about group propensities.

Course description

The students had completed a course that has been taught for many years by Allen Gagnon at Holyoke High School in Holyoke, MA. For the most part, the students who take the course are college-bound seniors. Beginning in 1991, Gagnon's class served as a test site for the development of the data analysis tool *DataScope*® (described below) and the accompanying lab materials developed by Konold and Miller (1994). Prior to this, Gagnon had taught a fairly traditional course using the text *Elementary Statistics* (Triola, 1986). At the same time that *DataScope* was being developed, Gagnon began teaching statistics more in the spirit of Tukey's (1977) Exploratory Data Analysis (EDA) and having students analyze real data using the computer. He now uses a text primarily to introduce various statistical techniques and concepts and spends the bulk of class time having students explore and discuss various datasets. Student activity is often structured around written labs that introduce the datasets, related aspects of the software, and proposing questions for exploration. In the course taught during the 1994 -1995 school year, from which the interviewed students were selected, class time had been split approximately evenly between probability and statistics. For the statistics component, students were introduced to various concepts including the median, mean, standard deviation, interquartile range, stem-and-leaf plot, boxplot, bar graph, histogram, scatterplot, and frequency table. Statistical inference received little attention.

DataScope, which the students used during the class and again during the interview, is specifically designed as an educational, rather than professional, tool (see Konold, 1995). In *DataScope*, raw data are entered and stored in a row (case) by column (variable) table. Students select variables for analysis by giving variable designations to columns of the data table and then selecting display commands. Basic analysis capabilities include tables of descriptive statistics (including the minimum, maximum, mean, median, standard deviation,

first and third quartiles), frequency tables (one- and two-way), bar graphs (that plot numeric data as histograms), boxplots, and scatterplots. The number of analysis capabilities is kept to a minimum.

DataScope was intended for courses that put less emphasis on the mechanics of statistics and more on the role of the statistician as detective, exploring complex data in search of interesting patterns and relationships. With this modest "tool kit," students can explore relationships among any combination of quantitative and qualitative variables. A general feature that facilitates this is a "grouping" capability, in which any variable can be split according to levels of one or more "grouping" variables. For example, if a variable "Height" (the height of various individuals) is grouped by a variable "Sex" (the gender of those same people), a number of different displays can easily be obtained that contrast the heights of males and females, including side-by-side tables of descriptive statistics, two-way frequency tables, and separate histograms or boxplots displayed one over the other for easy visual comparison.

PURPOSE OF STUDY

The purpose of this study was to explore difficulties encountered by students conducting fairly rudimentary data analysis using the computer. Although there is a growing body of research on people's understanding of statistical concepts (for reviews, see Garfield & Ahlgren, 1988; Shaughnessy, 1992), little of this research has been conducted with students doing anything complicated enough to be considered data analysis, where they are choosing both questions to pursue and the analysis methods they will use. Rather, the focus has been on understanding concepts, such as the mean (Mokros & Russell, 1995; Pollatsek, Lima, & Well, 1981) and the law of large numbers (Fong, Krantz, & Nisbett, 1986; Well, Pollatsek, & Boyce, 1990), that figure centrally in traditional statistics courses but that move to the background in problem/project-oriented courses that emphasize exploratory data analysis. Finally, much of this research has studied students with little or no instruction in statistics. Although this research offers insight into the knowledge and prior concepts students bring to their first course, we also need to better understand the kinds of problems and the nature of the problems that emerge during, and persist throughout, instruction as students encounter new concepts, methodologies, representational systems, and forms of argument. Hancock, Kaput, and Goldsmith (1992) and Lehrer and Romberg (1996) followed students during prolonged periods of data-analysis instruction and then described performance in more complex data analysis tasks.

METHOD

Students

Four volunteers (two pairs) were interviewed and paid for their participation. The students in the study, all females, had just completed the year-long course on probability and statistics taught at Holyoke High School. They had performed during the year at about the class median, and both pairs had previously worked together in class.

Materials and procedure

The students were interviewed by the first author as they worked together in pairs using *DataScope*. The interview sessions lasted approximately 90 minutes. We explained to the students that we wanted to see "how students with some background in statistics make use of the computer to analyze a set of real data." We posed a

series of questions concerning a dataset the students had explored during the course, and about which each student had conducted and written up an analysis of two questions they themselves had formulated. The dataset contained information on 154 students in mathematics courses at the school during that year and a previous year. The 62 variables included information about age, gender, religion, job status, parents' education, stance on abortion, and time spent in a number of activities, including studying, TV viewing, and reading.

The interview consisted of three major phases. During the first phase, the students were asked to use *DataScope* to give a brief summary of the dataset in order to characterize the students included in the survey. In this phase, we were interested in seeing which information they would focus on, which plots and summaries they would elect to use, and how they would interpret them. In the second phase, each student was reminded of one of the questions she had explored as part of her class project and was asked to use the computer to show what she had found. In the final phase, we posed a question they had not investigated during the course, which was whether holding a part-time job affects school performance. We asked them to investigate the question and, if possible, arrive at a conclusion. During all phases, the interviewer was free to pursue issues as they arose by asking follow-up questions. There were a few occasions when the interviewer took on the role of instructor, reminding one pair, for example, of the meaning of various parts of the boxplot when it became clear the pair had forgotten.

The computer (a Macintosh PowerBook) was placed between the two students. Each of them had access to a separate mouse and could therefore easily take the initiative during the interview. However, in each of the groups one of the students tended to take the lead. The interviewer did not attempt to alter this dynamic, and in general addressed questions to both students and accepted answers from either. The interviews were videotaped using two cameras, one focused on the students and the other on the computer screen.

RESULTS AND DISCUSSION

For the purposes of analysis, we produced complete transcripts of the interviews and augmented these with computer screen shots that showed the plots generated by the students during the interview. Transcript excerpts we present here are labeled with paragraph numbers that indicate the location of statements in the interview, and letters that identify the speakers. The student pairs consisted of J and M in one interview, and P and R in the other. Interviewer statements are labeled I. Ellipses (…) indicate omitted portions of the transcript, pauses in speech are represented by dashes (--), and a long dash (—) indicates a discontinuation of thought.

The students' motivation for analyzing data no doubt plays a role in determining both the techniques used and their persistence in analysis. During the class and interviews, students were encouraged to raise and explore questions of personal interest to them, but nothing of import was affected by the answers to those questions; thus, one could argue that these students were not performing as they might if something more critical were at stake. What motivated these students during the interview is a matter of conjecture. Although most of the time they appeared engaged and thoughtful, it is nevertheless the case that they were exploring the data primarily because we had asked them to, which may be similar to classroom motivation.

Overview of analysis methods used by students

Although the transcripts include examples of both good and poor statistical reasoning, what is most striking to us is that when they were given two groups to compare, both pairs of students rarely used a statistically-appropriate method of comparison. We do not mean by this that they failed to use a statistical test to determine

whether an observed difference between two statistics was significant or even that they failed to realize the need for such a test; rather, they did not use values in their comparisons that we would ordinarily regard as valid statistical indicators (e.g., means, percents, medians). Given that comparing two groups is fundamental in statistics, and that these students had just completed a course that had as one of its major foci methods of making such comparisons, this aspect of the students' performance warrants explanation. We will argue that the techniques these students do use in comparing two groups suggest that they have not yet made the necessary transition from thinking about and comparing either the properties of individual cases or the sizes of groups of individuals with certain properties, to thinking about and comparing group propensities.

Although there is evidence that M and J thought in terms of propensities when considering the distribution of cases on single variables, they did not use comparable reasoning to compare the distributions of separate groups on these same variables. Table 1 shows the types of statistical displays each pair of students used to explore questions involving two variables. M and J explored three questions involving the relation between two categorical variables (e.g., Are males or females more likely to have a driver's license?). Although they did use two-way frequency tables, most of their conclusions were incorrectly formed by evaluating the difference between two of the frequency counts in the table. In their interview, R and P did not investigate any questions involving two categorical variables. However, they used, as did M and J, two-way frequency tables almost exclusively in their analyses. Their reliance on two-way tables may be partly due to the fact that two-way tables were introduced near the end of the course; thus, they may have been using the most recently-practiced technique.

In all, R and P investigated seven questions and M and J investigated two questions that involved comparing two groups on a numeric variable (e.g., Do those with a curfew tend to study more hours than those without a curfew?). During the course, they had primarily used grouped boxplots to explore such questions, comparing medians to decide whether the two (or more) groups differed on the variable of interest. Less frequently, they had generated "grouped" tables of descriptive statistics, which displayed summary values including means and medians for each group.

During the interview, both pairs also used two-way tables to explore the relation between two numeric variables (e.g., Is there a relation between hours spent watching TV and school grades?), producing massive and unwieldy displays. Although both pairs tried on more than one occasion to look at a scatterplot, neither group successfully generated one, either because one of the variables selected was categorical or because they had not specified the variables appropriately in the syntax of *DataScope.*

Table 1: Types of statistical displays students used to explore three types of questions

	Students M & J		**Students R & P**	
	Display Type			
Question Type	Frequency Table	Table of Descriptive Statistics	Frequency Table	Table of Descriptive Statistics
Categorical x Categorical	3	0	0	0
Numeric x Categorical	2	0	5	2
Numeric x Numeric	2	0	1	0

In a number of instances, the four students exhibited difficulties interpreting boxplots and histograms, which they usually generated only at the interviewer's prompting. Both pairs expressed a preference for two-way tables, pointing out that precise values were available in a frequency table when those values would have to be

estimated on boxplots and histograms. Thus, another reason they may have primarily used frequency tables during the interview was that they could interpret the basic display elements (cell and marginal frequencies) much more easily than in boxplots, bar graphs, and histograms. Gal (in press) reports that even third-grade students are quite good at "literal" readings of two-way tables.

However, this explanation is not entirely satisfactory, because even if they knew they poorly understood various components of the boxplot, they all could explain basically what a median was, identify the median in a boxplot, and obtain values of either the median or mean from the table of descriptive statistics. Thus, it seems feasible that they could have used grouped boxplots or grouped tables of descriptive statistics rather than two-way frequency tables to compare two groups on a numeric variable. As mentioned, we suspect that a major reason for their preference for frequency tables, and their corresponding indifference to other displays available in *DataScope*, stems from their not holding a propensity view (i.e., from not having the understanding necessary for comparing groups composed of variable elements). Without this view, they tended to fall back on the nonstatistical comparison methods mentioned above. We argue that the two-way table allows them to do this. We find support for this hypothesis, which we discuss below, in the particular way in which they used frequency tables as well as in how they attempted to interpret histograms and boxplots.

Comparing absolute rather than relative frequencies

During the interview, M and J generated three two-by-two tables involving categorical variables. In each instance, they initially used frequencies rather than percents to make a group comparison. For example, they generated Table 2 to compare males and females with respect to holding a driver's license. In their memory, few of the males in their class had driver's licenses, and there was some joking between the two of them during the interview in which the boys were characterized as content with having their parents chauffeur them around--being lazy or perhaps indifferent to the privileges and prestige associated with having a license.

Table 2: Frequency table for "License" grouped by "Sex"

	License		
Sex	no	yes	total
f	35 (0.44)	45 (0.56)	80
m	19 (0.26)	54 (0.74)	73
total	54 (0.35)	99 (0.65)	153

Interview:

183. I: [Referring to the frequency table] So what is this, what does this tell you?
184. J: It shows that more males have licenses than females.
185 . I: And, so what numbers, what numbers are you looking at when you compare those?
186. M: Well, 54.
187. J: To 45 or ---
188. M: 54 out of the 73 boys that were interviewed have licenses, and 45 out of the 80 girls that were interviewed have licenses.
189. I: And, so which is ---
190. M: Well, the boys, more males have their license than females

Whether M in 188 was just reading off all the numbers in the table, a tendency we see elsewhere in the interviews, or was trying to communicate something about the rate of licenses is not clear. We think the former, since the focus in the rest of the protocol is clearly on the *number* of males and females with licenses. Their expectation seemed to be that *more* of the girls would have licenses than boys. And their final interpretation in 190 was worded in terms of absolute frequencies — "more males…"

In 191 below, the interviewer raised the possibility that because there were different numbers of boys and girls in the survey, the comparison of actual frequencies might not be valid.

I91 There are more females overall in the sample, right?… how do you take that into account in making comparisons.

192. M: I don't know.

193. I: Can you, can you do that if there are different numbers?

194. M: What?

195. I: Can you compre -- How do you deal with the fact that there are different numbers of males and females in trying to decide -- whether males are more likely to have licenses or not?

196. M: Well, you could look at the percentage, too. --- I guess.

197. I: Does that take care of it?

198. M: Well, I guess. I don't know.

199. J: I don't know how you would do it.

200. M: I guess, maybe. I'm not really clear on what you're asking.

201. I: Well, maybe I'll ask it again.…

Although M seemed unsure of the nature of the criticism, after a little thought she interrupted to offer an explanation.

203. I: I just want to ask you what —

204. M: What? Because there's a different number of males and females, the total?

205. I: Right.

206. M: Well, you could look at the percentage because the percentage would look at it differently. It would take the percentage of how many there are in total. So, it would take the percentage of that so, you could look at 74 percent of males have their license and 56 percent of females do.

207. I: So which way do you think is better, to report the absolute numbers or the percentages?

207. M: Probably, the percentages because there are more females than males so, you know, take the percentage of each would probably be more accurate, I guess.

M showed some awareness that percents permit comparisons of different sample sizes, but the "I guess" tagged to her final assertion suggests lingering doubts. Furthermore, later in the interview M and J resorted to comparing frequencies when interpreting another two-way table that cross-classified whether or not the students had (1) a driver's license, and (2) a curfew imposed on them by their parents. This case was particularly dramatic because different conclusions would be reached depending on whether one compared absolute or relative frequencies. They eventually did make the switch to comparing percents without prompting, but still demonstrated some tentativeness about whether and why percents are more appropriate than frequencies.

We argued above that simply comparing the size of two groups in which the group elements all share the same feature is not a statistical comparison because it says nothing of the propensity of those group elements to take on certain characteristics. To make the argument that this is what these students were doing, however, we have to establish that they were indeed attending only to one feature of the groups of males and females — in this case that they were interested only in the number of students of each gender who have licenses and were not somehow taking into consideration those without licenses.

Attending to values rather than dimensions

The students' statements associated with making group comparisons using frequency tables indicates that once they had used one variable to form discrete groups, they did not regard the other variable on which these groups would be compared as a variable at all. That is, they did not acknowledge one of the variables as a dimension along which different cases could fall. They attended instead only to a subset of cases with a specific value. For example, early in the interview M and J generated Table 3 to test their theory that more females held jobs than males.

Table 3: Frequency table for variables "Job" grouped by "Sex"

	Job		
Sex	no	yes	total
f	23 (0.29)	57 (0.71)	80
m	16 (0.22)	57 (0.78)	73
total	39 (0.25)	114 (0.75)	153

92. M: So, oh no, it's pretty even.
93. I: So tell me how — So, you're, you're looking at the percentage of males and —
94. M: Yeah.
95. J: Yeah, the difference.
96. M: Yeah, of females and males who have jobs.
97. J: Or who don't.
98. M: And that don't. And for the amount of--that do have jobs, that females and males are pretty even.
99. I: So what — tell me the numbers. I can't read them.
100. M: 57 males and 57 females.

We interpret M's correction in 98 of J's *or* (in 97) to *and* as signaling that she sees these as two separate questions. In 98 she also makes it clear that she is looking at those who hold jobs and in 100 uses only the students with jobs in her comparison. Indeed, if comparisons are going to be made on the basis of absolute numbers, this distinction is important. Using the frequencies in Table 3, they might well conclude that gender makes no difference with regard to holding a job, but does make a difference to not holding one. This might explain in part why, when asked to summarize such a table, R and P tended to read off all of the values.

Later in the interview M and J pursued the question of whether "people with licenses have a curfew." They compared the groups who did and did not have licenses only with respect to having a curfew; those who did not have curfews were not mentioned. After they had clearly indicated the values in the table (see Table 4) on which their comparison was based, the interviewer asked:

355. I: Okay and again, in that last comparison, we sort of ignored these two numbers [pointing to "no" curfew column]. Is that, is that all right to do?
356. M: Yeah, for what we are comparing.
357. I: And, how come?
358. J and M: We were looking at
359. J: license with [curfew], and not without. And you could do another question without, because our main thing was the license.
360. M: And the curfew.
360. I: Okay. So, if I asked you, "Are you less likely to have a curfew if you have a license vs. if you don't have a license?" then you'd look at those other numbers?
362. J and M: Yeah.

Table 4: Frequency table for "Curfew" grouped by "License"

	Curfew		
License	no	yes	total
no	17 (0.31)	37 (0.69)	54
yes	34 (0.35)	64 (0.65)	98
total	51 (0.34)	101 (0.66)	152

As mentioned above, earlier in the interview M and J had compared absolute frequencies rather than proportions. In this instance, they compared percentages rather than absolute frequencies. In the case of the two-by-two table, reporting that "the percentage of students with a curfew is 69% for those with a license and 65% for those without a license" is no different than reporting the complements of those percentages (31% and 35%) for the students without a curfew. But the above excerpt shows that M and J were not aware that in using percents they were now using all the information in the table.

The use of percentage had been suggested earlier in the interview (see paragraph 191) when the interviewer questioned the fairness of making comparisons based on different sample sizes. In our own statistics courses, we often introduce the need for relative frequencies in just this way. However, motivating the use of percents on the grounds of fairness may not bring with it an understanding of propensities. In this case, although M and J used percents rather than group size when prompted with the fairness issue, they did not appear to realize that as a result they were using a measure of propensity that reflected students both with and without a curfew. They still appear to be forming two groups (those with and without licenses) that do not vary on the other critical attribute (having a curfew) and then comparing the size (now measured in percents) of the groups.

In their study, Hancock et al. (1992) found that middle-school students preferred Venn diagrams to plots such as bar graphs that organize all cases along a dimension of values. They speculated that features that make these "axis" plots useful for detecting patterns and trends also make them more difficult to understand:

> Students find it easier to think in terms of qualities of objects ("this circle contains all the people who prefer rap") rather than spaces of possible qualities associated with a datafield ("this axis arranges people according to their music preference"). (p. 355)

Our findings suggest that one reason it is easier to think about attributes of objects as opposed to attribute spaces, or dimensions, is that in focusing on attributes one can circumvent the issue of variability. Once there is no variability in collections of values, one can use nonstatistical methods of comparison. Additionally, we found that even when students used dimensional plots such as two-way frequency tables and bar graphs, they tended to view and interpret them much like Hancock et al.'s (1992) subjects interpreted simple Venn diagrams: In their analyses, they still isolated elements that shared common values or attributes. Examples of this tendency are provided below.

USING FREQUENCY TABLES TO COMPARE TWO GROUPS ON A NUMERIC VARIABLE

In investigating the question of whether those with curfews studied more hours per week than those without curfews, R and P generated a 2 (Curfew) x 21 (Hours) frequency table (part of this table is shown in Table 5). Omitted columns are indicated by breaks in the table. Note that in producing frequency tables, *DataScope* does not classify numeric data into intervals as it does with histograms. This limitation can result in rather unwieldy displays.

Table 5: A portion of the frequency table for variable "Homework" gruped by "Curfew"

	Homework					
Curfew	0	12	14	15	27	total
no	7 (0.14)	1 (0.02)	2 (0.04)	4 (0.08)	1 (0.02)	50
yes	2 (0.02)	3 (0.03)	5 (0.05)	5 (0.05)	0 (0.00)	100
total	9 (0.06)	4 (0.03)	7 (0.05)	9 (0.06)	1 (0.01)	150

273: P: What was your question again?

274: R: If having a curfew affects your studying, like you study more if you have a curfew.

275: P: Well, I'm looking at like, 12 hours you get 3 people, and then 5, 5, you know, more people study more hours if they have a curfew.

276: R: But, there's also more people.

277: P: But, I mean it's like less people who don't [have a curfew]. You know, there's like 1 for 12 hours, 2 for 14. Do you know what I mean?

278: R: Yeah.

P focused in on a specific part of the table; that is, for specific hours of study (12 to 15) she compared the number of students with a curfew to those without a curfew. We are not sure why she chose to look at those particular values. But she seems to have assumed that the range 12-15 represents significant study time so that when she noticed that the numbers across this range were larger for students with a curfew than without, she concluded that those with a curfew were studying more. The basic technique she used involves isolating similar values of study time and then counting the numbers of students from each group at those values. There was no explicit attempt made to use the dimension; that is, to look for a trend as study hours increase. (See Biehler's chapter in this volume for further analysis of this difficulty).

In spite of R's expression of concern about the difference in numbers between those with and without curfews, R seemed to be convinced by P's argument. However, when the interviewer raised the issue again, R

reaffirmed her concern, pointing out that 100 students had curfews compared to 50 who did not. Asked if there was another way to deal with the problem of different group sizes, R suggested comparing "probabilities" because they "don't really have anything to do with how many people you have," and then used the proportional values in the table to do so. To demonstrate the difference between using probabilities and frequencies, she pointed out that in the case of 15 hours, one would draw different conclusions depending on whether one compared the 5 to the 4, or the .05 to the .08. Even though R uses the term "probabilities," which is highly suggestive of a propensity interpretation, we suspect she is not thinking about propensities, but only about fairness. Unable to distill from the table an overall conclusion, R finally suggested that they might make the judgment by selecting one value of study time, a value that seemed an ideal amount, and comparing curfew and no curfew groups at that value, effectively giving up all but a few of the 150 data points to form two groups in which there was no variability in the number of hours studied.

323. R: You could look, I mean, I would conclude that you would have to kind of pick a number to look look at. Like, say this is my limit. I say, if you study this many hours, you're going to have good grades, you're going to do good and it's just like if I pick 15 hours, I say that's good, that's how much time I think a person should go up to, should study up to, they shouldn't study more or less and then you would compare the two of them.

Why did the students in this interview revert to frequency tables whenever they could? R said that she, in general, preferred frequency tables to bar graphs. She wondered why one would bother trying to determine the exact height of each bar "when you can just look at the table and there you got it." M similarly expressed a preference for the frequency table "because you could just see the percentages and the numbers right there." The understandable preference for avoiding estimating when precise values are available does not explain why they did not make use of means and medians for comparing groups that were available in the tables of descriptive statistics.

Two features of frequency tables might explain their preference. First, as one moves from tables of frequencies to histograms and boxplots (in the case of numeric data), one can no longer identify either individual cases or the specific values they take on. From a statistical perspective, this is as it should be because what becomes important, and increasingly visible in histograms and boxplots, are group features. We think that what is especially important to these students about frequency tables in *DataScope* is that although the tables do not allow one to identify specific cases, they do the next "best" thing: They tally the number of cases at each *specific* value of the variable.

Having reached the impasse implied in paragraph 323, the interviewer asked R and P whether other plots might help them make the comparison. They first generated a display containing grouped boxplots, which showed separate boxplots of homework hours for each group, displayed one on top of the other over a common axis. Then they produced grouped histograms, again showing separate histograms of study hours for the two groups. To accomplish this using the software, they could leave the variable designations unchanged and simply activate the appropriate plot command. They expressed dissatisfaction with both of these plots, however, and attempted to use a feature of the software on both graphs to determine the number of students at particular levels of the variable homework hours.

367. I: What is it that you want to find out [about the boxplot display]?
368. R: Like how many students in 10 hours, you know like, how many students studied 10 hours on no [students without a curfew], and how many students studied 10 hours on yes [students with a curfew].

The software feature they wanted to use (a point identifier) does not operate on histograms, nor does it directly give the information they wanted for the boxplots. Notice, however, that they were trying to get from both these plots information they already had in the frequency table−−the number of cases for each group at each level of the dependent variable. There is some sense to this−−relating aspects of a well-understood representation to those of a poorly-understood one is a good way to improve one's understanding. Given that they subsequently showed confusion about interpreting both these types of graphs, this may have been part of their motivation. However, we also believe the students were trying to impose, on all three representations, a general method of making comparisons that makes sense to them−−that of looking for groups differences at corresponding levels of the dependent variable. Below, M and J showed the same tendency to isolate subgroups with grouped bar graphs, which they were using to determine whether fewer males than females had licenses:

246. I: You, you were just looking at these two columns [males and females with licenses] and ignoring those [males and females without licenses]. How come? Can you do that?
247. M: Well, because, well, the only reason that we were was because we were more interested in the people that had a license than the people that didn't. I mean, I guess it could, you could look at that, too, but —

In looking separately at each level of the variable, they were basically composing groups of elements all of the same type, and then simply comparing the number (and sometimes percent) of such elements in each group. We speculate that both pairs of students prefer frequency tables because they, more clearly than the other representations, segment values into discrete groups composed of elements that all have the same value.

EVIDENCE FOR A PROPENSITY INTERPRETATION

As mentioned above, there was some evidence in both interviews that the students were at times thinking in terms of propensities. The clearest examples were in the interview with M and J who used percentages on several occasions to summarize the distribution of a single, qualitative variable. These cases suggest that it was not their discomfort with, nor ignorance of, percentages that prevented them from using percents to compare values in the two-way table. Two of these instances of using percentages for a single variable immediately preceded their failure to use percentages in a two-way table. This raises an interesting question about why someone might have, but not make use of, a propensity perspective.

M and J produced Table 6 to show the marital status of students' parents.

Table 6: Frequency table for variable "Parents"

Parents	count	proportion
deceased	7	0.05
separated	43	0.28
together	103	0.67
total	153	1.00

152. M: So, most people's parents are together.

153. J: 28 percent are separated though.
154. I: You think that's fairly common
155. J: I think nowadays, yeah.
156. I: or representative of the rest of the school?
157. J: Oh, no.
158. M: I don't know. To have, I mean like it seems like 67 percent seems like a really high percentage to have parents together because with today, you know, so I don't know how that would have to do with the rest of the school. It would probably be a little bit less I think.

In evaluating the .67, M apparently compared the observed proportion to her own expectation, which was that the rate of divorce among families at her school was higher than what this sample suggested. M did not simply read values off the table, but offered a qualitative summary−−most people's parents are together." This episode demonstrates sound statistical practice in that M selected the most relevant values in the table, then interpreted and related to them to her expectations. We suggest that here the students were not just reading percents off the table, but were thinking in terms of propensities. That is, they were attending to the *rate* of divorce in the sample and not simply its frequency. This is indicated not only by the use of percents, but also by the use of the term "most" in 152, which relates the number divorced to the size of the entire sample.

It is instructive to think about what the implications would be if they had instead reported the frequency and tried to relate that to their expectations. To do so might have made sense if their expectations were something of the form "there are a lot of divorces these days." 43 divorces may, to some, seem like a large number regardless of the size of the comparison group. However, most people's expectations of the frequencies of divorce are probably more akin to ratios or percents given how often we hear that "half of all marriages end in divorce." In fact, it is hard to imagine how this and similar expectations could be mentally encoded as anything other than propensities, such that we would be able to compare a particular sample of a certain size to our expectations. We certainly do not store an entire series of expectations that tell us how many divorced people to expect in samples of varying sizes. Hancock et al. (1992) make a similar point when they puzzle over what the 8th-grade students in their study could have meant by their question "Can girls scream louder than boys?" if they were not thinking in terms of group means. Yet these students seemed prepared to decide the issue by comparing the totals of the individual loudness scores for the two (unequal-sized) groups.

M and J showed a similar pattern of responses with numeric variables, using appropriate summaries when thinking about the distribution of a single variable but inappropriate ones when comparing two groups with respect to that variable. For example, before looking at a distribution showing the number of hours worked per week, M explained the need to summarize the data.

417. M: …We could look at the mean of the hours they worked, or the median…because…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.

However, in trying to decide whether holding a job has a negative influence on school grades, they struggled with trying to interpret an eight (Grades) x two (Job) frequency table rather than comparing the mean or median grades of those who worked to those who did not. They attempted to determine whether there was a difference in grades by comparing the frequencies of those who did and did not work at each of the eight values of Grade (where the lower numbers represented better school performance), apparently expecting any effect to be evident in each row of the table. Although they did seem to be looking for a trend as they examined the grade values,

their conclusion was based on looking only at the results for those who received the top grades--the grades they said they would want to obtain.

We return to the question posed above: why did M and J stop using percents, and even regard percents somewhat skeptically, when they moved from examining the distribution of a single variable to examining the distribution of that same variable over two subgroups? One possible explanation is that in the one-variable case, M and J had generated an expectation to which the information reported in the table could be compared. As previously mentioned, it is difficult to imagine how such expectations could be encoded mentally as absolute numbers rather than as some form of propensity. However, in the two-variable case, there is no need to generate an expectation of the two values. The hypothesis M and J were investigating involved observing whether there was a difference between two values (fewer males than females have their licenses). The frequencies in the table seem to be perfectly adequate for deciding whether or not there is a difference. More concretely, when one examines a one-way frequency table showing the numbers of students having a license and sees the value 99, it is hard to do anything with that information in isolation; thus, 65% (or 99 out of 153) provides a context. However, when examining the two-way table and finding that 45 females versus 54 males have licenses, it might be easy to think that with those two values one has all that is needed for making a comparison. If this is a valid explanation of students' reasoning, we expect that if the task were slightly modified these students would use percentages.

For example, suppose the students were first asked to compare the instances of divorce in two groups (e.g., Catholics vs. Protestants) to the students' expectations of the overall divorce rate. If the students were then asked about the difference between the divorce rate of Catholics and Protestants, they might well compare percentages rather than absolute numbers.

A second possibility is that comparing two groups prompts a form of causal thinking based on reasoning about individual cases. In the single variable case, one does not need to think about particular causes to come up with an estimate. To answer the question "What's the current rate of divorce?", one need only have access to data, not to any information about what causal factors might be driving it. However, a question about divorce rates in two subgroups prompts one to wonder what might cause divorces to be higher in one group than in another. However, many students may not think in terms of what drives a *rate* but instead what drives *individuals* to divorce (Biehler, 1995). This is no longer a question of propensities, but of specific scenarios that lead to divorce. On several occasions during the interview, as the students were trying to explain why they thought groups might differ on some variable, they tended to focus on a single case (often based on their own experience) and to describe the details involved in that instance with no attempt to then talk about how more general causal factors might reveal themselves in a diverse group.

Both these explanations seem plausible, but we have not reanalyzed the interviews in an attempt to support or repudiate them. A third possible explanation, which we think is unlikely, is that the students' use of percentages in the single variable case has to do with the way in which the tables display the information. That is, it is easier to read the proportion in the one-way table than it is in the two-way table. Although *DataScope* does display proportions more prominently in the one-way table, this would not explain why percentages were not quickly adopted without question once they were pointed out.

CONCLUSION

To summarize, both pairs of students−−after a year-long course in which they had used a number of statistics including means, medians and percents to make group comparisons−−did not, without prompting, make use of these methods during the interview. We found evidence in the protocols that suggests that this failure was due in part to their having not made the transition from thinking about and comparing properties of individual cases, or properties of collections of homogeneous cases, to thinking about and comparing group propensities. Both pairs of students relied on frequency tables even when these were not the most appropriate displays. When using the frequency tables, the students tended to compare absolute rather than relative frequencies, even when groups differed dramatically in size. Group differences were judged by isolating those cases in the comparison groups that had the same value; thus, they effectively treated the dependent variable as if it were not a variable at all.

There is a large literature documenting peoples' difficulties interpreting two-way tables, and one could view our study simply as a further demonstration of these difficulties. However, most of the research concerning interpretation of two-way tables has explored people's ability to judge *association* (or statistical independence) by asking them whether the data suggest a *relationship* between the two variables, or whether one variable *depends* on the other. Note that the tasks in this study were generally construed as determining whether there was a *difference* between groups. Although judging group differences is formally comparable to judging association in these particular cases, these two types of judgments probably describe different cognitive tasks. For example, a common finding in the literature on judgments of association is that many people judge whether there is a relationship between, for example, having bronchial disease and smoking, by considering only one cell in the two-by-two table: the number of people who both smoke and have bronchial disease (Batanero, Estepa, Godino, & Green, 1996). This type of error, which we did not observe in our study, seems very unlikely to occur when the question posed is whether those who smoke are more likely than those who do not smoke to get bronchial disease.

We are not prepared at this point to offer specific prescriptions for how to foster the development and application of a propensity perspective. However, during analysis of these interviews we became aware of a general point we think has important educational implications; namely, the need to remember that the methods we use to compare groups are dependent on our reasons for comparing them. Consider, for example, the task that Gal, Rothschild, and Wagner (1989) presented to 3rd-graders and 6th-graders. They asked them to determine which of two groups of frogs jumped the farthest. If the purpose in making this judgment is to declare which group won the jumping contest, it is entirely adequate when the groups are of equal size to compare the totals of the individual scores of each group. Similarly, if they wanted to know whether a particular group of boys or girls can scream the loudest (Hancock et al., 1992), why not just total their individual performances? If the groups happen not to be equal in size, they might consider a variety of other options for how to compare them. In the "contest" framework, however, it makes sense for students to select criteria based on the fairness of those criteria; there is no requirement, however, that these also need to be measures of propensity. The advantage of using a measure of propensity for each group of jumping frogs or screaming students does not become apparent until we decide we would also like to know approximately how far frogs jump, or how loud boys and girls can scream. The students in our study may not have used propensities in the comparison tasks because they were reasoning from a contest framework. Defining the task for students as one not only of deciding, for example, whether those with curfews study more than those without curfews, but also of determining measures of how much those groups study, may have induced a more statistical approach.

Of course, our ultimate purpose is not to find ways to "induce" statistical reasoning, but to help students understand the relations among various purposes, questions, and methods so that they have more conscious control over the methods they select. This idea fits with recommendations recently made by Biehler (1995) who analyzed the apparent tensions between traditional, probability-based statistics and the model-free approach of exploratory data analysis (EDA). He points out that various practices in EDA (identifying outliers, breaking collectives into ever smaller groups in search of explainable differences) might encourage reasoning about individual cases at the expense of aggregate reasoning. His suggested resolution, however, eschews the idea that we teach strictly from one perspective or the other, but that we encourage students to explore relationships among various apparent antagonisms (e.g., between aggregate-based and individual-based reasoning or between deterministic and nondeterministic perspectives). In this approach, students are not asked to abandon explaining individual behavior, but rather to explore its power and limitations across various situations and in comparison to other perspectives.

Acknowledgments

We thank Rolf Biehler, Maxine Pfannkuch, Amy Robinson, Heinz Steinbring, and an anonymous reviewer for their helpful comments on earlier versions of this manuscript. This research was supported with funding from the National Science Foundation (RED-9452917) and done in collaboration with Rolf Biehler and Heinz Steinbring, University of Bielefeld, who conducted independent analyses of the same interviews. The opinions expressed here, however, are our own and not necessarily those of NSF or of our collaborators.

REFERENCES

Batanero, C., Estepa, A., Godino, J. D., & Green, D. R. (1996). Intuitive strategies and preconceptions about association in contingency tables. *Journal for Research in Mathematics Education*, *27*(2), 151-169.

Biehler, R. (1995). 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, Marrakech 1994*. Minneapolis: University of Minnesota.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology*, *18*, 253-292.

Gal, I. (in press). Assessing statistical knowledge as it relates to students' interpretation of data. In S. Lajoie (Ed.), *Reflections on statistics: Agendas for learning, teaching, and assessment in K-12*. Hillsdale, NJ: Erlbaum.

Gal, I., Rothschild, K., & Wagner, D. A. (1989). *Which group is better? The development of statistical reasoning in elementary school children*. Paper presented at the meeting of the Society for Research in Child Development, Kansas City, MO.

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education, 19*, 44-63.

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

Kaput, J., & West, M. (1993). Assessing proportion problem complexity. In G. Harel & J. Confrey (Eds.), *The development of multiplicative reasoning in the learning of mathematics*. *Research in Mathematics Education Series.* Albany: State University of New York Press.

Konold, C. (1995). Datenanalyse mit einfachen, didaktisch gestalteten Softwarewerkzeugen für Schülerinnen und Schüler [Designing data analysis tools for students]. *Computer und Unterricht, 17*, 42-49.

Konold, C., & Miller, C. (1994). *DataScope®*. Santa Barbara, CA: Intellimation Library for the Macintosh.

Lehrer, R., & Romberg, T. (1996). Exploring children's data modeling. *Cognition and Instruction, 14*(1), 69-108.

Mokros, J., & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education, 26(*1), 20-39.

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

Popper, K. R. (1982). *Quantum theory and the schism in physics*. Totowa, NJ: Rowman and Littlefield.

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

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

Shaughnessy, J. M. (1992). Research in probability and statistics: Reflections and directions. In D. Grouws (Eds.), *Handbook of research on the teaching and learning of mathematics* (pp. 465-494). New York: Macmillan.

Triola, M. F. (1986). *Elementary statistics* (3rd ed.). Menlo Park, CA: Benjamin/Cummings.

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

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