Statistics Education Research Journal

Volume 3 Number 2 November 2004

International Association for Statistical Education
http://www.stat.auckland.ac.nz/~iase

International Statistical Institute
http://www.cbs.nl/isi

Statistics Education Research Journal

SERJ is a peer-reviewed electronic journal of the International Association for Statistical Education (IASE) and the International Statistical Institute (ISI). SERJ is published twice a year and is free.

SERJ aims to advance research-based knowledge that can help to improve the teaching, learning, and understanding of statistics or probability at all educational levels and in both formal (classroom-based) and informal (out-of-classroom) contexts. Such research may examine, for example, cognitive, motivational, attitudinal, curricular, teaching-related, technology-related, organizational, or societal factors and processes that are related to the development and understanding of stochastic knowledge. In addition, research may focus on how people use or apply statistical and probabilistic information and ideas, broadly viewed.

The Journal encourages the submission of quality papers related to the above goals, such as reports of original research (both quantitative and qualitative), integrative and critical reviews of research literature, analyses of research-based theoretical and methodological models, and other types of papers described in full in the Guidelines for Authors. All papers are reviewed internally by an Associate Editor or Editor, and are blind-reviewed by at least two external referees. Contributions in English are recommended. Contributions in French and Spanish will also be considered. A submitted paper must not have been published before or be under consideration for publication elsewhere.

Further information and guidelines for authors are available at: http://www.stat.auckland.ac.nz/serj/

Submissions

Manuscripts should be sent to co-editor Flavia Jolliffe (F.Jolliffe@kent.ac.uk), by email, as an attached document in Word format. These files should be produced using the Template available online. Full details regarding submission are given in the Guidelines for Authors on the Journal's Web page: http://www.stat.auckland.ac.nz/serj

Publication: IASE/ISI, Voorgurg, The Netherlands
Technical Production: University of New England, Armidale, NSW, Australia

ISSN: 1570-1824

SERJ EDITORIAL BOARD

TABLE OF CONTENTS

EDITORIAL

This issue marks the end of SERJ's third year of operation. It is special in several ways, both because of its content and of the fact that we are announcing several important changes in SERJ's Goals and Policy Statement as well as in the type of papers we publish. We thus encourage you to read further, and follow forthcoming updates to our guidelines on the SERJ website: www.stat.auckland.ac.nz/serj.

Focus: Variation. This issue is a Special issue focused on research on reasoning about variation and variability. We thank the Guest Editors, Joan Garfield (University of Minnesota, USA) and Dani Ben-Zvi (University of Haifa, Israel) for organizing and supporting this project. Their Forward paper describes the background of this Special issue and introduces the five papers in it, of which one is invited and four are refereed. We are also planning to have a special section with several more papers on the same topic of variation and variability as part of our next issue planned for May 2005. Joan Garfield and Dani Ben-Zvi will again serve as Guest Editors of this special section. In this way we hope to extend the contribution of new research and reflective papers to current knowledge in the important yet little-studied area of reasoning about variation.

Revised goals for SERJ. Recently, SERJ's Editorial Board has adopted a new Goals and Policy Statement which expands in several ways the scope of issues and topics sought in manuscripts. Below is our revised statement, followed by brief explanations of our rationale.

> SERJ aims to advance research-based knowledge that can help to improve the teaching, learning, and understanding of statistics or probability at all educational levels and in both formal (classroom-based) and informal (out-of-classroom) contexts. Such research may examine, for example, cognitive, motivational, attitudinal, curricular, teaching-related, technology-based, organizational, or societal factors and processes that are related to the development and understanding of stochastic knowledge. In addition, research may focus on how people use or apply statistical and probabilistic information and ideas, broadly viewed.
>
> The Journal encourages the submission of quality papers related to the above goals, such as reports of original research (both quantitative and qualitative), integrative and critical reviews of research literature, analyses of research-based theoretical and methodological models, and other types of papers described in full in the Guidelines for Authors. All papers are reviewed internally by an Associate Editor or Editor, and are blind-reviewed by at least two external referees. Contributions in English are recommended. Contributions in French and Spanish will also be considered. A submitted paper must not have been published before or be under consideration for publication elsewhere.

This statement maintains SERJ's focus on research related to core aspects of learning and teaching of statistics and probability in classroom-based contexts at primary, secondary, and tertiary levels. However, it reflects the growing recognition that learning occurs in many types of out-of-school contexts, such as in the workplace or at home. People of all walks of life, not only "pupils" or "students", engage with diverse tasks and situations where knowledge of statistics or probability, however acquired, is called for and put to use. The growing availability of computers, and the dissemination of statistical data and information via the Internet, further contribute to blurring of traditional boundaries between formal and informal learning. Hence, teaching, learning, and using are increasingly intertwined and have to be seen as occurring within a broadening social sphere. Our revised Goals and Policy Statement therefore encourages research that addresses an expanded set of issues related to improving the way people understand and use statistical and probabilistic knowledge.

Paper types and revised author guidelines. The updated Goals and Policy Statement lists a range of the kinds of paper which are sought by SERJ. In addition to research papers based on diverse methodologies, we are also encouraging other types. Among these are reflective or theoretical analyses, epistemological studies, and critical or integrative literature reviews which are based on research and which can contribute to future research, theory-building, or educational practice.

We are also pleased to announce **a new paper format - Brief Reports**. Such papers, up to 2500 words, can report, for example, on replication and extension studies, psychometric studies, results of program evaluations, or preliminary conclusions from innovative research projects. While such studies could of course lead to full-length manuscripts, the possibility of submitting a Brief Report offers international researchers an additional and more economical publication channel that has the potential for faster turnaround, while maintaining the same scientific standards as in full-length papers.

We are presently revising our Guidelines for Authors, in light of the changes described here, as well as due to the need to fine-tune details of paper formatting and other technical aspects of preparation of papers either for submission for review or for publication. Further details about Brief Reports and about changes in paper submissions will be included in these revised Guidelines for Authors which will appear in early 2005 on the SERJ Website. Please contact either of the editors if you seek information on the Brief Report format before the updated Guidelines for Authors are published.

Alert regarding duplicate conference and journal submissions. We draw readers' attention to issues concerning submissions of papers originally written for a conference for possible publication in SERJ. Many authors use a conference presentation as a springboard for preparation of a paper for later submission to a research journal; we are happy to be part of that cycle. However, due to the blurring of what a "publication" means in this age where the Internet enables rapid worldwide availability of full papers, we want to reiterate the need to avoid duplicate publication, which is a standard policy in many journals. Our policy is that papers "published" by conference organizers (on the Internet, in printed Proceedings, or on CD) will *not* be accepted for consideration and review by SERJ unless they include substantial new data and textual material, beyond what appeared in the published conference paper. This applies to both refereed and non-refereed conference papers.

In closing, we thank our readers for their continued interest and support of SERJ, which is reflected in the increasing number of entries to and downloads from the SERJ website. We are especially grateful to our dedicated referees, whose names are listed in this issue. Finally, we thank the other members of SERJ's Editorial Board, whose counsel and support underlies the announcements of the various changes announced in this issue.

FLAVIA JOLLIFFE AND IDDO GAL

RESEARCH ON REASONING ABOUT VARIABILITY: A FORWARD

GUEST EDITORS:

DANI BEN-ZVI
University of Haifa, Israel
dbenzvi@univ.haifa.ac.il

JOAN GARFIELD
University of Minnesota, USA
jbg@umn.edu

We are very pleased to introduce this special issue of the Statistics Education Research Journal (SERJ), which presents cutting-edge research in an area of increasing importance: *Reasoning about variability*. The notion of variability and the importance of its role in statistics have been documented by David Moore, the well-known statistician and former president of IASE and ASA. Moore (1990) describes statistical thinking as recognizing the omnipresence of variability and considering appropriate ways to quantify and model the variability of data. Wild and Pfannkuch (1999) further describe the centrality of variability in statistical thinking, as revealed by their studies of expert statisticians solving statistical problems: "*Variation is the reason why people have had to develop sophisticated statistical methods to filter out any messages in data from the surrounding noise*" (p. 236).

We note that the terms *variability* and *variation* are often used in the teaching and research literature interchangeably, and this may add to a confusion regarding this complex area. Reading and Shaughnessy (2004) address this problem and offer the following definitions for the two terms: "*Variation* is a noun used to describe the act of varying or changing condition, and *variability* is one noun form of the adjective *variable,* meaning that something is apt or liable to vary or change" (p. 201). However, they note that educators and researchers often refer to *variability* as the characteristic of the entity that is observable, and the term *variation* as the measuring of that characteristic. So a distinction is made between what is observed (what is varying) and what is measured. Moore (1997) points out that both variability and the measuring and modeling of that variability are important, however he does not distinguish between terms used to describe these phenomena. The papers in this special issue do not necessarily follow these definitions; however they mostly refer to "variability" to represent how data vary.

Despite the attention paid by statisticians and statistics instructors to this important topic, to date little has been published about how people, particularly novices and statistics students, actually reason about variability, or how this type of reasoning develops in the course of statistics instruction. Examples for challenging questions that call for careful attention by researchers and educators are: What are the simplest forms of variability that children can understand? What are instructional tasks and technological tools that promote the understanding of variability? What are the common misconceptions regarding variability? What are the difficulties that people encounter when dealing with variability in data? What does correct reasoning about variability look like? What are ways to assess understanding of variability? How does an understanding of variability connect and effect understanding of other statistical concepts and types of reasoning? What are useful methodologies for studying the understanding of variability? What type of understanding of variability is sufficient for a statistically literate person?

Statistics Education Research Journal 3(2), 4-6, http://www.stat.auckland.ac.nz/serj

In response to these challenges, "*Reasoning about Variability*" became the theme of the third and most recent international research forum on Statistical Reasoning, Thinking and Literacy (SRTL-3), held in July 2003 at the University of Nebraska, USA. The SRTL-3 forum built on the two previous SRTL forums, held in 1999 in Israel and in 2001 in Australia. The five papers appearing in this special issue include one invited paper (Gould) and four peer-refereed papers (Hammerman & Rubin, Ben-Zvi, Bakker, Reading) that are based in part on presentations and discussions at SRTL-3.

A unique aspect of the SRTL forums is the opportunity to bring together a small number of researchers whose work is focused in a particular area. They have an opportunity to present their research in extended sessions that permit lengthy discussions among the participants. In addition, many researchers present video clips of students discussing and explaining their actions and reasoning; this allows for intensive review and discussion of research methods and results by all participants.

At SRTL-3, a variety of researchers from diverse backgrounds and countries presented studies that examined different aspects of reasoning about variability. For example, some looked at the inherent variability of data, variability as represented in a univariate or bivariate distributions, the role of variability in comparing groups, students' understanding of particular measures of variability (e.g., the standard deviation), and variability in different sampling contexts (Lee, 2003). The studies presented involved students from elementary school through college, and some studies also examined teachers' reasoning about variability. After five days of presentations and discussions, the participants were reinforced in their belief that variability is a complex topic to understand, learn and teach, and that its understanding is a fundamental component of statistical reasoning and thinking. We also began to design a hypothetical learning trajectory that may be useful to help guide teachers as they aim to develop students' understanding of this topic. We see the role of innovative technological tools, appropriate teacher guidance and curricular tasks, as well as using good data sets, as crucial in helping students develop this idea.

The first paper in this Special Issue, by *Gould*, is based on his opening address at SRTL-3 and designed to provide a statistician's view of the importance of noticing, understanding and analyzing variability and using measures of variation when making sense of data. This paper frames issues and provides examples that help to appreciate the complexity of the conceptual issues that researchers and educators have to grapple with. The paper by *Hammerman & Rubin* focuses on secondary level teachers who are learning statistics and attempting to cope with variation in data using the innovative Tinkerplots software (Konold & Miller, 2004). The paper by *Ben-Zvi* provides a detailed qualitative analysis of the ways by which two seventh grade students started to develop views (and tools to support them) of variability in comparing groups task using various statistical representations. The paper by *Bakker* describes activities and use of technological tools by advanced 8th-grade students and how their reasoning about variability and variation is developing. The paper by *Reading* suggests hierarchies to assess high school students' understanding of variation, one for more qualitative descriptions and the other for more quantitative descriptions of variation.

The research studies presented in this special issue have several common features. Their topics reflect the shift in emphasis in statistics instruction, from statistical techniques, formulas, and procedures to developing statistical reasoning and thinking. These studies employ various types of qualitative methodologies, which appear to have uncovered many interesting points about how students and teachers reason about variability. Most of them use extended teaching experiments, or represent cases where researchers collaborated with teachers in field settings or designed specialized learning episodes or environments, to be able to elicit detailed and deep data about students' actions and reasoning.

Most of the studies in this special issue emphasize the role of technology (statistical software or specially designed tools) in developing students' statistical reasoning about the variability of data. This is not surprising, given how the discipline of statistics has depended on technology and how technology has been driving changes in statistical practice. Although there are many technological tools available, including graphing calculators, computers, and the World Wide Web, there is still a lack of research on how to best use these tools and how they affect student learning. Regardless of the type of technology used or the level of the students studied, all the studies attempt to understand the

development of conceptual models that students (or teachers) use to reason about data and its variability. Together, these studies help us understand the complexity of the idea of variability, and its interconnectedness to core statistical ideas of data, sampling, distribution, and center.

The forthcoming issue of SERJ (May 2005) will also include a special section of papers related to reasoning about variability which will enable us to continue the attention to research on this important area. Among them will be two invited papers which will reflect on the collection of papers in this special issue, by *Cliff Konold* (University of Massachusetts, Amherst, USA) and *Maxine Pfannkuch* (University of Auckland, New Zealand). We hope that their responses will lead to productive discussions about the importance of the notion of variability in statistics education as well as about ways to further study and improve its development in students at different educational levels and contexts.

We appreciate having the opportunity to put together these papers and responses for publication in SERJ, and especially value all the contributions of the co-editors, in particular, *Iddo Gal* (University of Haifa, Israel), who oversaw this special issue and offered many suggestions to improve the quality of the papers. We also thank all of the participants at SRTL-3 who contributed to discussions of earlier versions of these papers and to those who served among the reviewers of these papers. We invite readers with comments and suggestions to contact us. Finally, we invite researchers to review information (see "Forthcoming Conferences" in this issue) on the forthcoming SRTL-4, to be held in 2005 in New Zealand, which will be devoted to *Reasoning about Distributions*.

REFERENCES

Konold, C., & Miller, C. (2004). *Tinkerplots*, version 0.93. Data analysis software for the middle school. Amherst: University of Massachusetts.

Lee, C. (Ed.) (2003). *Proceedings of the Third International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-3)*. [CD-Rom, with video segments] Mount Pleasant, Michigan: Central Michigan University.
[Online: http://www.cst.cmich.edu/users/lee1c/SRTL3/]

Moore, D. (1990). Uncertainty. In L. A. Steen (Ed.), *On the Shoulders of Giants: New Approaches to Numeracy* (pp. 95–137). Washington: National Academy Press.

Moore, D. (1997). New pedagogy and new content: The case for statistics. *International Statistical Review, 65,* 123–165.

Reading, C., & Shaughnessy, J. M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literary, Reasoning, and Thinking* (pp. 201–226). Dordrecht, The Netherlands: Kluwer Academic Publishers.

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

DANI BEN-ZVI
Computers in Education Program, Head
Faculty of Education
University of Haifa
Mount Carmel, Haifa 31905, Israel

JOAN GARFIELD
Department of Educational Psychology
315 Burton Hall
178 Pillsbury Drive, S.E.
University of Minnesota
Minneapolis, MN 55455, USA

VARIABILITY: ONE STATISTICIAN'S VIEW

ROBERT GOULD
Department of Statistics, UCLA
rgould@stat.ucla.edu

SUMMARY

Although variability is of fundamental concern and interest to statisticians, often this does not get communicated to students who are taught instead to view variability as a nuisance parameter. A brief survey of a few case studies, as well as a recounting of some history, shows that variability is worthy of study in its own right, and examination of variability leads to insights that might have been missed had we focused all of our attention on the "trend" of the data. As one of the key components of statistical thinking, variability deserves more prominence in the classroom.

Keywords: Variability; Education; Statistical thinking

1. INTRODUCTION

Many of the papers in this special issue discuss naive conceptions of variation. Much has been written on more experienced conceptions, particularly in the context of statistical thinking. Moore (1990), for example, puts variation at the heart of the process of statistical thinking and addresses the needs of statistical thinkers to acknowledge the omnipresence of variation, to consider variation in collecting data, to quantify variation, and to explain variation. Wild and Pfannkuch (1999) provide a thorough overview on the topic, placing variation in the context of a rather rich model of statistical thinking. When educators think about how to move students from their naive conceptions towards a more "professional" view, an understanding about how practitioners confront variation should be useful, and it is hoped that this paper provides some examples that will aid in the process. Statisticians are themselves a variable crew and these remarks should not be taken as a summary of the Profession, but instead as the thoughts of one practitioner.

It is fair to say that statisticians have a complex relationship with variability. Statisticians sometimes attempt to minimize variability, sometimes to maximize, sometimes to estimate or simply to "analyze" variance. Many statistics educators claim variability to be one of the fundamental concepts of statistics, for example, Moore (1990) and Snee (1990). Yet when most students first encounter statistics, they find that variability plays second fiddle to "central tendency". The conceptualization of data as "signal versus noise", which according to Pfannkuch (1997) some statisticians consider one of the major contributions of Statistics to Science, teaches students that the central tendency, however it's measured, is of primary importance and variability is simply a nuisance. A noisy one at that.

College level statistics does not completely ignore variability, of course. Many texts and one hopes many instructors discuss the importance of examining the shape of the distribution before making any conclusions about the data. DeVeux, Velleman and Bock (2004) write in their introductory statistics textbook that "the three rules of data analysis are 1) make a picture 2) make a picture and 3) make a picture." Most students learn, often in the first weeks of the course, that the mean by itself is not a sufficient summary of a distribution. But after that variability is brushed aside as attention focuses on estimating the mean, and students are taught that standard deviation is a nuisance parameter that must be estimated if one is to do a proper hypothesis test on the mean or

Statistics Education Research Journal 3(2), 7-16, http://www.stat.auckland.ac.nz/serj

calculate a confidence interval for the mean or perform a comparison of means. Some introductory courses teach ANOVA, which although it pays tribute to variability in its name, is really about the simultaneous comparison of means from several populations.

The definition for variability used in this discussion is derived from Moore's (1997) definition of data analysis as "the examination of patterns and striking deviations from those patterns". Although Moore was describing the activity of data analysis as a by-product, he provides a wonderfully general definition of variation; variation is that which is not pattern. A case study in Section 2 will illustrate how useful and rich variability, using this broad definition, can be. A short examination of the history of statistics in Section 3 shows that Variation and Center have a long-standing rivalry. Finally, Section 4 illustrates the role variability plays in the analysis of three small data sets.

2. AN EXAMPLE OF REASONING WITH VARIABILITY

In 1982, Morton et. al. conducted a study to determine whether parents who worked around lead could expose their children to dangerous amounts of lead. Lead poisoning is particularly dangerous for children because excess levels of lead interfere with a child's development. For example, lead-based paint is no longer used in the interior of homes because children might ingest flakes of paint. But adults who work around lead might get sufficient amounts of lead dust on their skin or clothing to pose a hazard to their children.

The data consist of the blood lead levels (measured in micrograms per deciliter) of 33 children whose parents worked at a battery factory in Oklahoma and whose work exposed them to lead. We also have the lead levels for a control group consisting of an additional 33 children, matched for age and neighborhood, whose parents did not work around lead. The data themselves are taken from Trumbo (2001).

The "research question" is phrased so as to invite a comparison of means. Is the typical lead level of the exposed children higher than the typical lead level of the control children? It is instructive to imagine a world in which means or medians are computed only as a last resort, and attempt to answer this question without relying on the concept of central tendency. Consider the histograms of the lead levels of the two groups of children (Fig. 1) below, which are presented with only frequency information. One displays the distribution of the control group, the other the exposed group. Can you tell which is which?

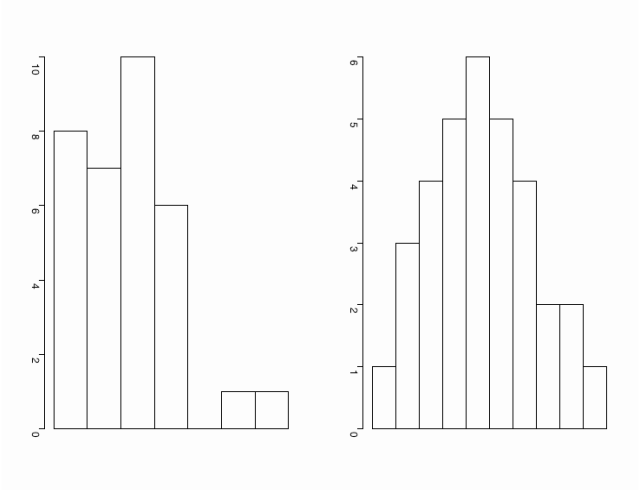

Figure 1. Lead levels in blood of 33 children. Which histogram comes from the exposed group?

Even though one does not know the scale (actual values on the x-axis) of each group, nor their means or spread, it is fairly straightforward to identify the right-skewed histogram (on the left) as belonging to the exposed group. The reason is that the shape of the sample distribution is consistent with our theory of how the exposure happens; children receive varying amounts of additional

exposure beyond the "normal" exposure represented by the control group. Thus, shape, even in the absence of information about absolute values, contributes important information that helps us in making sense of the data in light of the research question.

Of course, knowing the actual values (shown in Figure 2) and through them the means of the two groups contributes additional understanding. Indeed, the numeric values themselves carry quite a bit of the story. Most medical experts find lead levels of 40 mg/dl and above to be hazardous, and levels of 60 and above to require immediate hospitalization. Without any significance testing, it's quite apparent that the means are different and that the exposed group is dangerously higher.

At its heart, though, this study is concerned with a causal question: did the parents' exposure at work cause their children's elevated levels? While the difference in means establishes that the causal question is worth entertaining, it provides little evidence towards answering the question. Because this is an observational study, not a controlled experiment, definitive causal conclusions are impossible. However, I propose that the difference in shape between the two groups, while by no means confirming causality, by itself takes us closer to a causal conclusion than would a consideration of the means alone. The reason for this is that the proposed mechanism for the children's contamination had consequences for both the center and the shape of the distribution of lead levels. Had the centers differed but the shapes been similar we would not have been inclined to believe that the parents' exposure could have been the cause, but might instead suspected, perhaps, deliberate poisoning. On the other hand, had the centers for both groups been the same but the shapes appeared as they do here, then we would still have had strong reason for suspecting that parents' exposure to lead was a threat to children.

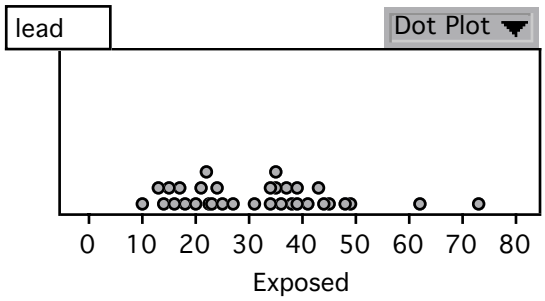

Figure 2. Blood lead levels for both groups of children

The interplay between the center and the spread of distributions of data is, of course, not new. A brief survey of the history of modern statistics will show that there has long been some sort of "tension" among data analysts as to the proper roles for center and spread.

3. HISTORICAL OVERVIEW

I make no claims to being a historian, and I use no primary sources for this survey. Unless otherwise noted, the 19th century history is a paraphrase of Gigerenzer et al. (1997) and the history of ANOVA is from Searle, Casella, and McCulloch (1992).

Adolphe Quetelet, impressed by growing evidence of "statistical regularities" in practically every aspect of life that governments measured in the early 1800's, founded "social physics" to determine laws that governed society that would be analogous to those laws that governed the motion of the planets. Some examples of statistical regularities, such as the fairly constant ratio between male and female births, were well known at this time and not too surprising. But other regularities were more surprising because they seemed to suggest an implicit order arising from chaos. Examples of such regularities include Laplace's demonstration that the number of dead letters at the Paris postal system was fairly constant, as well as those later "discovered" in the homicide, crime, and marriage rates published in the 1827 volume of the French judicial statistics.

Quetelet, along with others, believed that these regularities were not just descriptions of societies, but instead represented some underlying "reality" about society. He invented *l'homme moyen*, the

Average Man, to be an abstraction of the typical member of a society. While there might not be laws that governed individuals, there were laws, Quetelet believed, that governed the behavior of the Average Man. Hence the Average Man was not a mere description of a society, but something more real.

Quetelet's views were influential throughout Europe for much of the 19th century. Florence Nightengale corresponded with him and called him the founder of "the most important science in the whole world" (Coen, 1984). Adolph Wagner, a German economist, believed the power of statistical regularity was so strong that in 1864 he compared it to a ruler with power so great that it could decree how many suicides, murders, and crimes there would be each year in the kingdom. But, of course, this ruler lacked the power to predict precisely who would commit, or fall victim to, these acts.

The Average Man was a useful tool because he could be used to compare traits of different cultures and, presumably, his behavior (or more accurately, his propensity towards certain types of behavior) could be predicted. However, he had troubling implications for the concept of free will. If there were a quota for murders, by what mechanism were people compelled to fill it? Surely people could choose to not fill the quota, if they wished. Interestingly for our purposes, by arguing in support of the existence of free will in society, critics of the Average Man also argued in support of elevating the use of variability in statistical analyses of social data.

Gustav Rumelin claimed that variability was a characteristic of the higher life forms and reflected autonomy. Humans have free will and are therefore more variable, presumably, than single-celled organisms. If humans were homogenous, then Statistics would be unnecessary, and therefore social statistics should study variation rather than simply reporting the average. Wilhelm Lexis, a German social statistician, studied the annual dispersions of these so-called statistical regularities and compared them to chance models. In almost every case he found that the observed dispersions were greater than that predicted by his chance models, and used this to conclude that the existence of free-will prevented the existence of statistical regularity and, therefore, studying averages of populations was a waste of time.

A generation later, in 1925, R.A. Fisher invented ANOVA to cover the need for "a more exact analysis of the causes of human variability." Ironically, ANOVA really tends to treat variability as a nuisance and its main focus, once one is satisfied that the variance is behaving, is to concentrate on comparing means. Nonetheless, once ANOVA was later framed in the context of the linear model, it became possible for researchers to model and investigate variance components directly.

While this historical overview is a selective review, it is clear that discussions of the interplay between measures of variability and measures of center have long been a part of modern statistical development. Social scientists' struggle to understand how, when, and whether to assess variability has been complex and at times controversial, but valuable. Data analysts have learned that one must consider both the center and the variability in order to understand scientific and social phenomena.

4. CASE STUDIES

The following three case studies should illustrate how and what data analysts learn from variation. While the data presented here are real, the analyses are not. The analyses are meant to be instructive and are not necessarily what a data analyst would actually perform.

4.1. CASE STUDY 1: UCLA RAIN

Time-series analyses are notoriously signal-noise oriented and provide a good opportunity to examine the role of variability in a context where one might think that role is secondary. This series is monthly rainfall at the campus of the University of California, Los Angeles from January 1936 to the end of June 2003. A typical goal of an analysis of data of this type might be to model the trend so that predictions for the future could be made, but here we focus more on how our view of the trend changes as we re-organize the data.

The three graphs below show, respectively, the time-series with the overall average superimposed (Figure 3), rainfalls organized by month (Figure 4) and a smoothed time-series showing total annual rainfall (Figure 5).

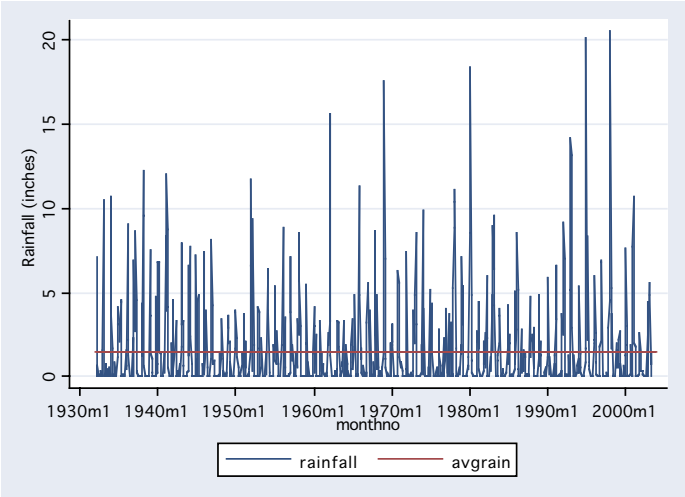

Figure 3. Rainfall in inches at UCLA

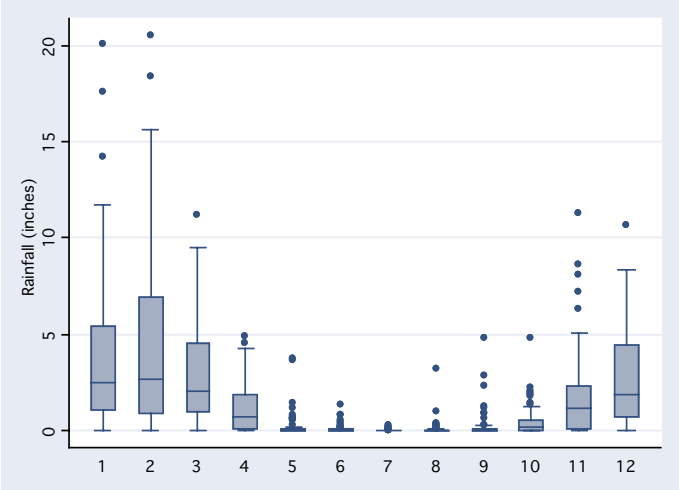

Figure 4. Rainfall by month

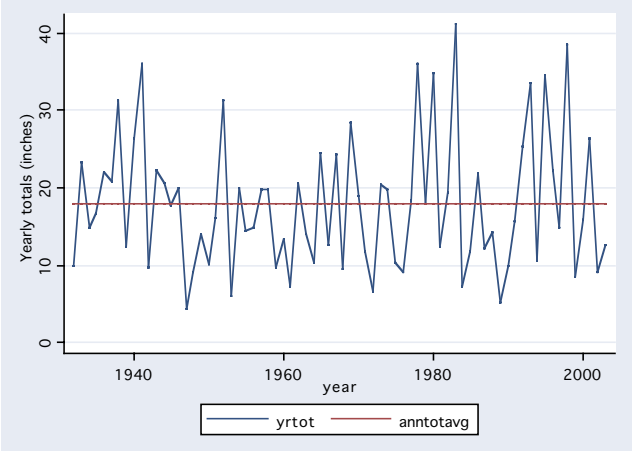

Figure 5. "Smoothed" total annual rainfall, with average annual rainfall indicated by horizontal line

The unprocessed time-series (Figure 3) impresses mainly by its unruliness. The second graph (Figure 4) shows a pattern any southern Californian would recognize: wet winters, dry summers. The third graph (Figure 5) shows an historic trend and one can, for example, search for evidence of drought. These last two graphs illustrate the relationship between center and variation suggested within Moore's definition of data analysis. Variation is defined in contrast to pattern. In the second graph, we can see variation within a particular month; for example we can see that there was a particularly rainy September with almost 5 inches of rain. We also notice that this amount is substantial for any month, but is particularly heavy for September. We can also see variation across months which provides an understanding of seasonal fluctuation. The third graph shows us annual variation with respect to an overall mean, which might be of interest to farmers or climatologists.

4.2. CASE 2: LONGITUDINAL DRINKING PATTERNS

Do people drink less as they age? This was a question investigated by Moore et al. (2005). Drinking patterns are fairly complex in that drinking varies quite a bit from person to person, and individuals vary their drinking from year to year. One difficulty in such an analysis is in isolating the effects of cohort (people born at a certain historic time might share drinking patterns) and period (changes in price or supply might affect the entire population at a certain point in time.)

The data come from the 1971, '82, '87, and '92 waves of the NHANES study, a national, longitudinal random sample of about 18000 U.S. residents (NCHS 1973 10a, 1973 10b, 1987, 1990, 1992, and 1994). Subjects were asked questions about the quantity and frequency of their drinking, and responses were converted into a "quantity/frequency index" (qfi) that corresponds approximately to the average number of drinks per week. Figure 6 shows what 1971 looked like. The story, once again, is in the variety of drinking. Note the outlier above 80.

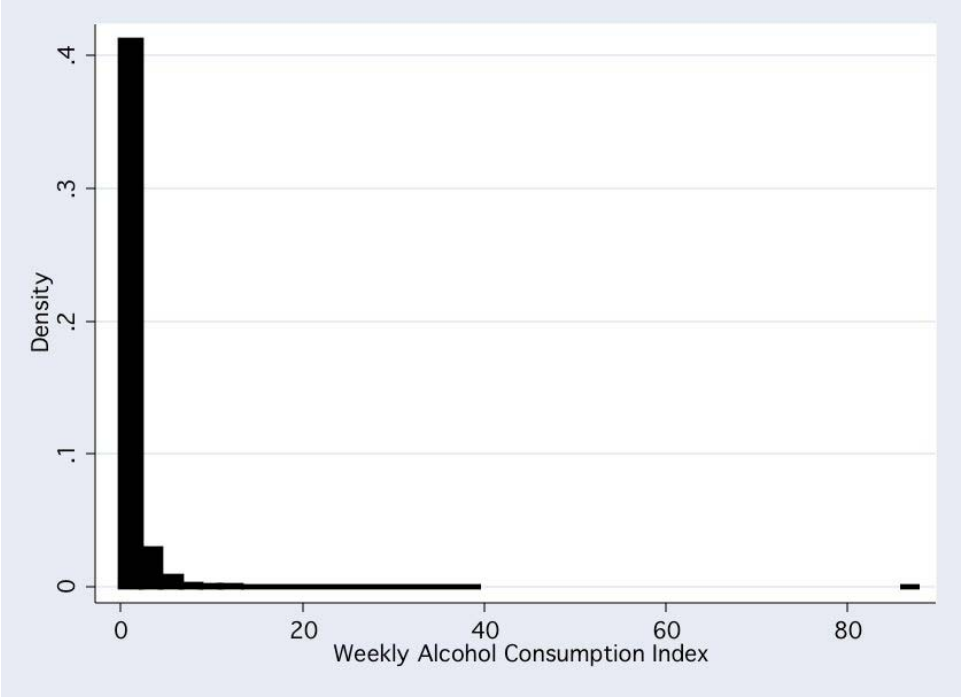

Figure 6. Drinks per week of a national sample in 1971

We see that the vast majority drinks little, but a minority drinks very much. The shape of the distribution is interesting in that it tells us that a simple model, in which we look for "typical" drinking with some people deviating from the norm, will be inappropriate. At the very least a log transformation of the data is necessary, and thus our consideration of variation has affected our conception of the model.

The purpose of this study was to examine drinking over time. Figure 7 shows a Log transformation of the QFI and indicates that drinking did in fact vary at the different waves of the survey.

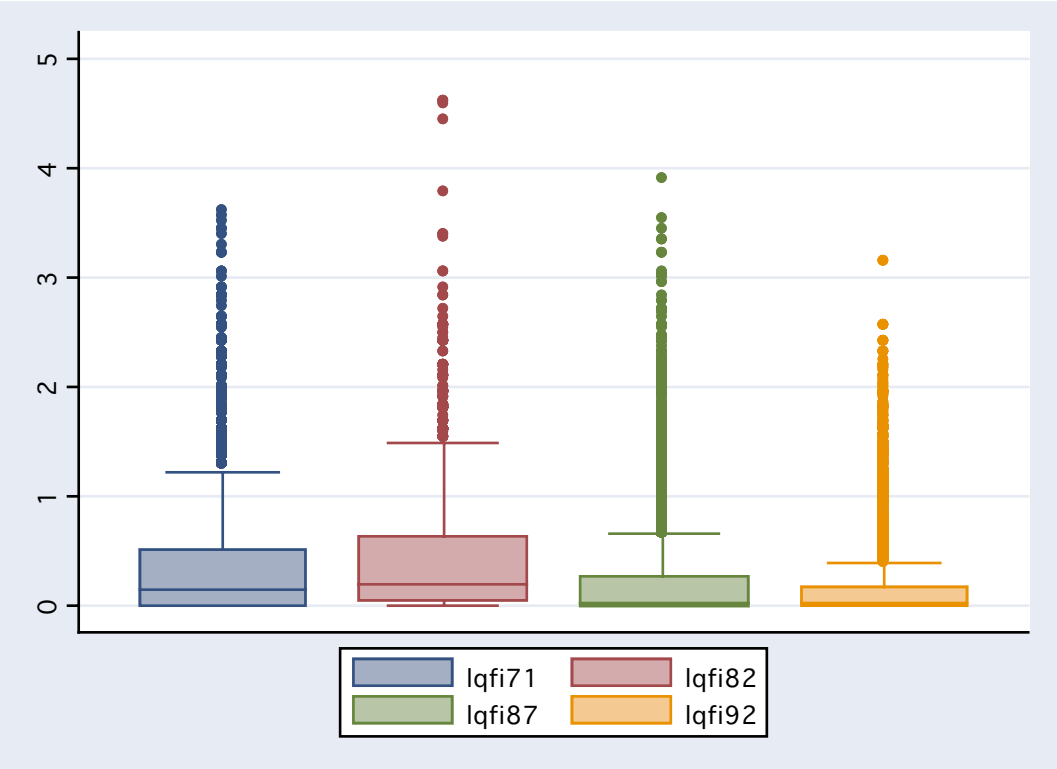

Figure 7. Drinking index at each wave of the survey

One is compelled, at this point, to attempt to explain the variation with some sort of model. As a first cut, many find it useful to talk about two types of variation: explained and unexplained, or if you prefer, deterministic and stochastic (Wild & Pfannkuck, 1997). Deterministic variation is that which we believe will have a regular structure, a structure that can be defined by a model. What is left over is then stochastic variation.

Let's consider a very simple model in which the only explanation for variation is time. We would then have log(qfi+1) = 0.97 - 0.017*year as one model for the deterministic variation; we could conclude that drinking amounts decreased slightly over time, while acknowledging that there was still quite a bit of variance left unexplained. A more complex model would then chip away at the unexplained variation bit by bit. For example, another source of variation is the individual; different people drink different amounts and will change differently over time. We could then fit a mixed linear model in which each individual is allowed his or her own slope and intercept (Laird & Ware, 1982). The variation is now much more complicated; we have variation with respect to each individual's path as well as variation between individuals. Examination of these different sources of variation might lead to further refinements of the model and force us to consider such questions as whether observations within individuals are independent and whether slopes are correlated with intercepts.

One potential deterministic model for these data that includes age, cohort, and period explanations for variation is log(qfi+1) = 0.4 - 0.13*(age in decades) + 0.18*(per capita consumption in alcohol) + .035*(birthyear times age in decades). This model, if valid, suggests that drinking declines with age, across all generations and periods of (recent) history, but the decline depends on when a person was born. Those born more recently decline more slowly. We used per capita alcohol consumption to control for historic variations in drinking. So this model says that even in times in which the country as a whole drank more (or less), individuals on the average declined as they aged.

Although the end result of this analysis is a model for the trend, the model has been shaped and refined by our conceptualization of the variation.

4.3. CASE 3: CHIPMUNKS

About 15000 years ago, during the Wisconsin glacial stage, chipmunks that lived in the pinyon-juniper woodland in the Mojave Desert region were able to move from mountain to mountain, since the cooler temperatures allowed their pinyon and juniper trees to grow in the basins between mountains. Later, these woodland areas retreated to higher elevations and with them the chipmunks. This resulted in isolated chipmunk communities. Kelly Thomas, a graduate student at UCLA, wanted to study morphological differences in separated chipmunk populations. She captured several chipmunks at six different sites, took five morphological measurements of each chipmunk, and wanted to compare them to see if there were differences in size and shape at different sites. This is a fairly common activity for population biologists. They seek to quantify the shape of animals using, ideally, a small set of numbers.

Principal Components Analysis (PCA) is a data reduction technique that focuses on the co-variation of multi-variable samples. We use it here to attempt to reduce the number of variables we'll use to compare the chipmunks from five down to two. PCA does this by creating a set of linear combinations of the original data that maximize the variation and are orthogonal to one another. The reason for maximizing the variance is so that the resulting set of measurements will have the greatest possible dispersion. This, in turn, will make it easier to distinguish between populations of chipmunks. This is analogous to writing an exam to distinguish those who learned the material from those who did not. If everyone receives approximately the same score, it is difficult to distinguish those who really understood the material.

In the following analysis, each chipmunk provided two scores. Each score was a different linear combination of its mass, body length, tail length, hind-foot size and ear length. The first score emphasized overall mass and the second corresponded roughly to shape. Although the procedure was not entirely successful with these data (the two scores accounted for only 50% of the total variation), we gained some insight into the data. First, ear measurements were not strongly correlated with any of the other measurements. On the other hand, chipmunks with long bodies tended to have long tails, and those with big hind feet tended to have the greatest mass. More interestingly, by plotting the two scores for all chipmunks, we were able to discern that chipmunks from the same sites tended to have similar scores, which provided evidence that these scores could be useful for distinguishing chipmunks from different regions.

This procedure is often used for exploratory analysis. This example used PCA to informally assess similarities among chipmunks in similar sites. Interestingly, we did this by dealing directly with the variation and covariation among the variables.

5. CONCLUSION

This paper has presented several examples illustrating how a statistician thinks about variability. It falls on educators to consider how conceptions of variation aid or hinder how students learn statistical thinking.

Wild and Pfannkuch (1999) mention imagination as one of eight "dispositions" that statistical thinkers possess. My belief is that variation is the fuel to statistical imagination. The case studies presented above illustrate, to various degrees, how consideration of variation drives the analysis by provoking the statistical imagination to explore alternative models.

Statistical imagination begins when variation is observed. When confronted with a time-series, one intuitively seeks for some sense of order out of the chaos. Might some of the "noise" be abated by removing seasonal effects? Monthly effects? We can aggregate the data in different ways to get different pictures of the "trend", but in all ways, we are exploring different models of the variation in order to better observe the trend.

By defining variation, we define trend. The process of modeling variation is made explicit in time-series analysis, in which analysts consider explicit structures for variation to take into account

local correlations between observations. Modeling variation and covariation is also done explicitly when analyzing longitudinal models. The alcohol consumption case study discusses implications of including different sources of variation in the model. For example, if one thinks (quite reasonably) that a source of variation is due to different drinking behaviors between individuals, then one has a more complex model than if one simply models the population as a homogeneous mass. Analysts must explicitly build other assumptions about variability directly into the model. How do subsequent observations within a person correlate? How do people within a subgroup co-vary with each other? How do basic parameters of the model -- for example the rate at which people change their alcohol consumption as they age -- correlate among individuals? The statistical imagination, guided by expert substantive opinion, shapes the model through consideration of variation.

Models play an important role in statistics, and one hopes that introductory statistics students, at least at the college level, learn not only to interpret basic statistical models, but also to develop an understanding of the sometimes tenuous relationship between the model and reality. There is however a danger, I believe, in over-emphasizing models to beginning students.

Chatfield (1988) distinguishes between "confirmatory" and "exploratory" analyses, and it is models used in confirmatory analyses that I think should be de-emphasized. Confirmatory models harden the boundaries between "signal" and "noise". The conceptualization of data as "signal and noise" is of course very important, but perhaps the value-laden language creates too much of a sense of finality and inhibits statistical imagination in students. *This* is signal, but *this* is noise, and one should not pay attention to noise. In practice, one *does* pay attention to statistical noise. Confirmatory models do provide the all-important p-value, but students who possess (or are taught to possess) statistical imagination will see that models can be used for exploration and insight without necessarily needing that final confirmation step. Principal components analysis is one example in which the variation is the primary focus of the analysis, and, at least in the chipmunk case study above, no model is produced or required. I don't mean to overstate my case here. There is a model, in a general sense, underlying the analysis. But it is used as just one step in an exploration, not as a final statement about the structure of the data or reality. The lead case study is a good example of how the "signal" by itself (or at least a narrowly defined signal) provides an incomplete analysis and, in fact, isn't sufficient for answering basic research questions concerning the effects of lead exposure on children.

An important reason for focusing students' attention on variation is to encourage them to think not in terms of procedures ("Which test do I apply here?") but instead to exercise their statistical imaginations in order to understand the real issues behind the data. Often this translates to a search for causes of variability. Statistics courses perhaps give short shrift to the problem of inferring causality and are known for merely cautioning students against making conclusions about causality based on association. But students want, and maybe even need, causal explanations.

Pearl (2000) makes the point that causal explanations make an early appearance in the Bible. "Did you eat that apple?" God asks Adam. "Eve made me" is Adam's answer (greatly paraphrased). Causality makes for good story-telling, and links data to reality. The search for causal explanations can lead to heightened statistical imagination, and so students should be given the opportunity to reason, for example, about how an intervention will affect the shape of a distribution. Assuming that exposure to lead does lead to higher lead levels, why are the shapes of the lead distributions in Figures 1 and 2 natural?

If our primary goal is to teach statistical thinking, rather than statistical techniques, then we should look to the noise, and not the signal.

ACKNOWLEDGEMENTS

Many thanks to the direction of Joan Garfield and encouragement of Dani Ben-Zvi. Special thanks to James Murakami, Department of Atmospheric Sciences, UCLA for the rain data. I would like to dedicate this paper to the memory of Winifred Adams Lyon.

REFERENCES

Chatfield, C. (1988). *Problem solving: A statistician's guide.* London: Chapman and Hall.

Coen, I. B. (1984). Florence Nightingale. *Scientific American 250*, 128–137.

DeVeaux, R., Velleman, P., & Bock, D. (2004). *Intro Stats*. New York: Addison Wesley.

Gigerenzer, G., Switjtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1997). *The empire of chance: How probability changed science and every day life.* Cambridge, UK: Cambridge University Press.

Laird, N. M., & Ware, H. (1982). Random Effects Models for Longitudinal Data. *Biometrics, 38,* 963–974.

Moore, A., Gould, R., Reuben, D., Greendale, G., Carter, K., Zhou, K., & Karlamangla, A. (2005). Do Adults Drink Less as They Age? Longitudinal Patterns of Alcohol Consumption in the U.S. Accepted for publication in *American Journal of Public Health*, March 2005.

Moore, D. (1997). Probability and statistics in the core curriculum. In J. Dossey (Ed.), *Confronting the core curriculum* (pp. 92–97). USA: Mathematical Association of America.

Moore, D. (1990). Uncertainty. In L. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95–138). Washington, DC: National Academy Press.

Morton, D. E., Saah, A. J., Silberg, S. L., Owens, W. L., Roberts, M. A., & Saah, M. D. (1982). Lead absorption in children of employees in a lead-related industry. *American Journal of Epidemiology, 115*, 549–555.

National Center for Health Statistics. Plan and operation of the National Health and Nutrition Examination Survey, United States. 1971-73. *Vital Health Stat [1].* 1973;10a. DHEW publication PHS 79-1310.

National Center for Health Statistics. Plan and operation of the National Health and Nutrition Examination Survey, United States. 1971-73. *Vital Health Stat [1].* 1973;10b. DHEW publication PHS 79-1310.

National Center for Health Statistics. Plan and operation of the NHANES I Epidemiologic Follow-up Study 1982-84. *Vital Health Stat [1]*. 1987;22. DHHS publication PHS 87-1324.

National Center for Health Statistics. Plan and operation of the NHANES I Epidemiologic Follow-up Study 1986. *Vital Health Stat [1]*. 1990;25. DHHS publication PHS 90-1307.

National Center for Health Statistics. Plan and operation of the NHANES I Epidemiologic Follow-up Study 1987. *Vital Health Stat [1]*. 1992;27. DHHS publication PHS 92-1303.

National Center for Health Statistics. Statistical Issues in Analyzing the NHANES I Epidemiologic Followup Study. *Vital and Health Statistics, Series 2: Data Evaluation and Methods Research No. 121*. Hyattsville MD: National Center for Health Statistics, 1994. DHHS Publication No. (PHS) 94–1395.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge, UK: Cambridge University Press.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: John Wiley & Sons.

Snee, R. D. (1990). Statistical thinking and its contribution to quality. *The American Statistician, 44*, 116–121.

Thomas, K. (2002). Effects of habitat fragmentation on montane small mammal populations in the Mojave National Preserve. Unpublished report for UCLA OBEE 297A, UCLA, USA, 2002.

Trumbo, B. (2001). *Learning statistics with real data*. San Francisco: Duxbury Press.

Wild, C. J., & Phannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review, 67*(3), 223–265.

ROBERT GOULD
Dept. of Statistics
8130 Math Science Building
MC 155404
UCLA
Los Angeles, CA 90095-1554, USA

STRATEGIES FOR MANAGING STATISTICAL COMPLEXITY WITH NEW SOFTWARE TOOLS

JAMES K. HAMMERMAN AND ANDEE RUBIN
TERC
Jim_Hammerman@terc.edu; Andee_Rubin@terc.edu

SUMMARY

New software tools for data analysis provide rich opportunities for representing and understanding data. However, little research has been done on how learners use these tools to think about data, nor how that affects teaching. This paper describes several ways that learners use new software tools to deal with variability in analyzing data, specifically in the context of comparing groups. The two methods we discuss are 1) reducing the apparent variability in a data set by grouping the values using numerical bins or cut points and 2) using proportions to interpret the relationship between bin size and group size. This work is based on our observations of middle- and high-school teachers in a professional development seminar, as well as of students in these teachers' classrooms, and in a 13-week sixth grade teaching experiment. We conclude with remarks on the implications of these uses of new software tools for research and teaching.

***Keywords*:** Representations; Software tools; Variability; Proportional reasoning; Group comparison; Covariation; "Binning"

1. OVERVIEW

This paper reports on research at the intersection of two lines of inquiry: 1) What approaches do people use to deal with variability in data? and 2) How do new statistical visualization tools change the strategies people use in analyzing data? Each of these questions is worthy of study in its own right. Variability, while at the heart of statistics, presents a significant challenge to teachers and students trying to develop a sophisticated set of statistical reasoning strategies. Its ubiquitous presence in data makes simple statements that take *all* the data into consideration impossible, unless one can somehow acknowledge and "tame" the variability by working with fewer numbers. In some ways, a single measure such as the mean is the ultimate way to deal with variability in a distribution, since its role is, in fact, to reflect all the values in a data set with just one number. However, values such as the mean are notoriously difficult for students and teachers to understand as representing the entire data set at once (Konold, Higgins, Russell, & Khalil, 2003; Konold, Pollatsek, Well, & Gagnon, 1997; Konold et al., 2002; Mokros & Russell, 1995). So there is a need for other ways to deal with variability that teachers and students can understand and appropriate.

Methods for handling the variability in data depend intimately on the tools at hand, and therefore new software visualization tools are dramatically changing the way data analysis and statistics are learned and taught. Until recently, and to some extent still, the techniques most students and their teachers learned for describing a data set were primarily numerical, i.e., computing measures of center such as the mean or median, and computing measures of variability such as the standard deviation or inter-quartile range (IQR). The five-number summaries illustrated by box plots went one step further in describing a distribution by including both central tendency and some indications of variability. All these numerical characterizations have been made easier to obtain by the accessibility of calculators, often with built-in statistical functions, so that any student is able to compute some basic statistics about a distribution, even if she/he doesn't really know what they mean. With these current tools, computations in general are easy, as are some rudimentary graphical manipulations, but the new kind

Statistics Education Research Journal 3(2), 17-41, http://www.stat.auckland.ac.nz/serj

of tool we discuss in this paper greatly expands the possible ways teachers and students can interact with data.

Using any new tool or representation necessitates change in the content and pedagogy of statistics instruction and in many cases teachers are unprepared for these changes. Even the primarily pencil and paper data displays developed to simplify the task of visualizing a distribution in the past 30 years (Tukey, 1977) can present a challenge to teachers. Dot plots, stem and leaf plots, and box plots are particularly popular in elementary and middle school textbooks, and classroom conversations about data are now expected to include more nuanced descriptions of distributions that include the shape of the data, including issues of center, variability, skewness, and other characteristics best uncovered with graphical representations. Many teachers, however, have limited experience with these new approaches, so the potential for deeper classroom discussion afforded by these representations may be lost. New interactive visualization tools, such as the one described in this paper, have the potential to support further changes in classroom approaches to data analysis, and thus present yet another challenge to teachers. To effectively teach students about data analysis, teachers will have to both know the mathematical content and understand students' statistical reasoning as it develops. Because of these new classroom responsibilities, we have worked primarily with teachers, as mediators of student learning.

1.1. THE POTENTIAL EFFECTS OF NEW INTERACTIVE VISUALIZATION TOOLS

New interactive visualization tools (e.g., Tabletop™ (Hancock, 1995), Fathom™ Dynamic Statistics™ (Key Curriculum Press, 2000), and TinkerPlots™ (Konold & Miller, 2004)) create yet another set of new possibilities both for data display and for attaching numbers to graphs in informative ways. No longer are we limited only to numbers that we can calculate from formulas, often without looking at the distribution itself, and to a few static graphs. People using these software tools easily see the shape of data and other distributional characteristics that numerical summaries alone can't show without constructing graphs by hand. More important, they can dynamically manipulate data displays, and begin to get a feel for the data themselves and for relationships among data attributes. These software tools make possible a pedagogical change that embodies a view of data analysis that allows and even encourages students to transform graphs interactively and to request a variety of numbers that represent the distribution as it is displayed in the graphs.

These new tools by themselves don't reduce the complexity of data analysis, nor do they solve the problem of variability. People still have to attend to what can be an overwhelming amount of information, i.e., all the data values, although visualization tools hope to take advantage of our built-in perceptual apparatus to ease this task. In addition, these tools give people more options for describing distributions in ways that can be useful in statistical decision-making. The affordances these new tools provide and how people use them have been explored to some extent, primarily by developers of the particular tools and their collaborators (Bakker & Gravemeijer, 2004; Cobb, 1999; Hancock, Kaput, & Goldsmith, 1992; Konold, 2002; Konold et al., 1997; Rubin, 1991, 2002; Rubin & Bruce, 1991). However, these are only preliminary studies and barely scratch the surface of the inherent flexibility and complexity of using these tools. Researchers and teachers need more detailed information about what happens when learners analyze data with this kind of software over an extended period of time.

Interactive software data visualization tools which allow for the creation of novel representations of data open up new possibilities for students (and teachers) to make sense of data, but also place new demands on teachers to assess the validity of the arguments that students are making with these representations, and to facilitate conversations in productive ways (cf. Ball, 1993; Ball, 2001). Just as teachers in general need a deeper understanding of the mathematics they teach to work effectively in reform-oriented classrooms (Ball, 1991; Ball, Hill, Rowan, & Schilling, 2002; Borko & Putnam, 1995; Hill & Ball, 2003; Russell et al., 1995; Schifter, 1997; Sowder, Philipp, Armstrong, & Schappelle, 1998), so, too, will teachers need a deeper understanding of data and statistics (Rubin & Rosebery, 1988; Russell et al., 2002) to use new software tools in their classrooms. Professional development for teachers will need to address issues of mathematical content, as well as issues of learning, representation, and pedagogy. By exploring and discussing data for themselves in new ways,

teachers can develop a deeper understanding of the mathematics, and also of how classroom discourse and pedagogy might change through use of new software tools. However, *teachers'* experiences of learning with these new software tools have not yet been explored.

Many of the teachers came to our professional development seminar thinking that statistics is about mean, median, and mode. They knew how to calculate these statistics, though they didn't always have robust images about what they meant or how they were used. In general, they had not dealt with data sets that required them to confront significant variability, so that they didn't have strategies to apply to the new complexity they encountered using interactive visualization tools to explore real data sets. Finding that there was a lot more to see and analyze in data than they had ever imagined was both powerful and intimidating. We were interested in how these teachers, with their diverse prior skills and experiences in data analysis, represented and made arguments about data. Although our work was primarily with teachers, the strategies and approaches we saw them use were similar to those students used in the classrooms we observed.

This paper, then, describes research that primarily involved teachers who were learning to analyze data in a professional development seminar as well as students in several middle school classes. Our research sought to answer two questions: 1) What statistical reasoning strategies do teachers employ to handle issues of variability when analysing data? and 2) What new affordances does a tool such as TinkerPlots™ provide for coping with variability? We will describe ways that students and teachers used the new options available to them in TinkerPlots™ (Konold & Miller, 2004) to compare groups. In this context, there is almost never a way to make a clear judgment by simply producing a picture or two, except in the rare instance when there is no overlap between the distributions to be compared. How does a choice of representation and measure help simplify the difficult task of making a decision about two distributions which each have significant variability? We will describe two main approaches that are made possible by TinkerPlots™—those that use categorizing or "binning" continuous data to reduce the apparent variability; and those that use proportional reasoning, primarily to deal with issues of unequal group sizes. We describe several examples of ways in which teachers and students with access to TinkerPlots™ use each approach, comment on the validity of each technique, demonstrate how each approach attempts to confront and handle some of the complexity and variability inherent in data, and comment on the developmental course of the use of some of these strategies.

1.2. A PERSPECTIVE ON LEARNING

Our view of learning influenced a variety of aspects of our research: our choice of format for our teacher seminars, our choice of data sets and discussion topics, and our conclusions about teachers' and students' reasoning. Because learning to analyze data is a process that unfolds over time, we believe that we need a series of rich tasks and deep discussions to understand teachers' approaches and how they develop. Although we offered a data analysis seminar for teachers, our focus was less on teaching them specific concepts or techniques than it was on providing an environment in which teachers could explore important ideas about data and statistics using new software tools, and on conducting research on their thinking. Our teaching was, thus, strongly constructivist and our teaching and research were closely integrated as we asked questions and facilitated discussions focusing on the different ways that teachers were making sense of problems (cf. Duckworth's (1996) notion of "Teaching as Research"). Teachers often shared the several ways they made sense of particular problems, and discussions served to clarify these different approaches through questioning and explanations. Not all disagreements were resolved immediately, but teachers seemed comfortable letting confusing ideas simmer and re-emerge over the course of several weeks.

From this perspective, it is also important that the data sets and problems we provided were complicated enough to make important issues in data analysis salient. For example, if the data sets that learners are given to compare always have the same number of cases in each group, they will never have to confront the issue of unequal group size, or to think about ways to mathematically "equalize" the numbers in each group—they'll get the same results whether they compare groups using absolute numbers or percentages. It is only when the groups are *different* sizes that the

difference between using numbers or percentages becomes relevant. More complicated data sets, messier ones, are both more realistic and can give learners a chance to grapple with some important statistical ideas, though increased complexity may sometimes overwhelm novice learners. Helping learners manage the confusion that comes with this complexity so that it doesn't overwhelm them is an important part of our teaching.

Finally, we believe that learning is a slow, non-linear process of constructing and building more and more robust understandings over time. Powerful ideas, especially those requiring a move towards abstraction, may need to be built and re-built several times before they become solid and can be used in a wide variety of situations. They may appear to be in place at one point in time, but later, can appear to be "lost" when they are needed in a different, more complex context. The example of sixth graders embracing and then questioning the use of percentages (described in section 4.2) illustrates this. This view of learning is an important lens to use in reading the descriptions that follow, as teachers and students may appear to be reasoning inconsistently. Iterative building and rebuilding of ideas, we claim, is one of the hallmarks of the learning of important and difficult concepts.

2. REVIEW OF THE LITERATURE

Some of what makes working with data complex is the tension between attending simultaneously to individual values and to aggregate properties of distributions. While expert data analysts move fluidly between appropriate attention to these different levels of understanding data, this is a difficult perspective to attain. Several other views of data typically precede seeing it as a distribution with aggregate properties. One common perspective is to view data as a collection of individual values without any relationship to one another (Bakker & Gravemeijer, 2004; Hancock et al., 1992).

Konold and colleagues (Konold & Higgins, 2002; Konold et al., 2003) argue that children see data in *several* simpler ways before ever noticing aggregate and emergent features of data sets. Their fourfold schema includes the following different ways of viewing data, which we consider useful for examining the thinking of adults as well as children:

1. Data as a *pointer* to the data collection event but without a focus on actual data values—in this view, data remind children of their experiences, "We looked at plants. It was fun."
2. Data as a focus on the identity of individual *cases*—these can be personally identifiable, "That's my plant! It's 18 cm tall," extreme values, "The tallest plant was 37 cm," or interesting in some other way.
3. Data as a *classifier* which focuses on frequencies of particular attribute values, or "slices," without an overall view—"There were more plants that were 15 to 20 cm than 10 to 15 cm."
4. Data as an *aggregate*, focusing on overall and emergent characteristics of the data set as a whole, for example, seeing it as describing variability around a center, or "noise" around an underlying "signal" (Konold & Pollatsek, 2002)—"These plants typically grow to between 15 and 20 cm."

It is possible to make group comparisons from any of the case, classifier, or aggregate perspectives, i.e., comparing extreme values, comparing the numbers in particular slices, or the more canonical comparison of means, respectively. However, aggregate views are preferable, as they are required to look beyond the data towards making inferences about the underlying populations or processes represented by data samples. Konold and Pollatsek (2002) argue that it is sometimes easier to use aggregate measures of center when comparing groups than when looking at data distributions on their own. When comparing groups, they claim, it is clear that the focus is on underlying processes rather than the particulars of the data at hand, although earlier work (Konold et al., 1997) found that students had difficulties thinking of underlying "propensities" even when comparing groups. In the work reported here, both teachers and students only rarely used formal measures of center to characterize data sets even when comparing groups. Still, among the several methods we will report, there were some that demonstrated aggregate thinking without involving the use of measures of center.

As we and others suggest, data are most interesting when they are used to make inferences beyond themselves, that is, when they are seen as representative of a larger population about which one wants to generalize. This process of generalizing from a sample to a population is notoriously difficult. When comparing groups with data seen as a sample, the inherent variability of a particular set of data is complicated further by the fact that we must also determine whether observed differences in data reflect underlying differences in populations, or are merely due to chance fluctuations in a sample. The difficulty that people have in understanding the relationship between sampling variability and the inherent variability of the underlying population is well documented (Rubin, Bruce, & Tenney, 1990; Saldanha & Thompson, 2002; Sedlmeier & Gigerenzer, 1997; Watson & Moritz, 2000). Yet, while issues of sampling variability sometimes arose for teachers and students in this study, such variability is not the focus of this paper. We will, however, point to instances when issues of sampling were salient.

The TinkerPlots™ software we used in this study made it easy to divide data into "bins" in which cases within a range of values of an attribute are grouped together. The impact of such a representation has been little explored. Cobb (1999) reports how students using his minitools (Cobb, Gravemeijer, Doorman, & Bowers, 1999) were able to partition data in equal sized groups, allowing them to make arguments about the position of the middle half of the data (using a representation akin to box plots); and were able to partition data into groups with a specified interval width, allowing for arguments about the numbers or percentages on either side of a fixed value. Cobb, McLain and Gravemeijer's (2003) 8th grade (age 13) study argues for the utility of breaking bivariate data into a series of distributions of equal width "slices" of the independent variable in order to look for patterns in the position of these distributions and their means across the slices, that is, splitting bivariate data into a series of group comparisons which, they argue, is conceptually (if not actually) what expert data analysts do when looking for a regression line. Meletiou and Lee (2002) describe difficulties that students have with histograms, another form of grouping data, stating that "the research literature tells us very little about how understanding of histograms and other graphical representations develops" and calling for further research. Finally, some studies argue that students often like to characterize data by "hills" (Cobb, 1999), "modal clumps" (Konold et al., 2002) or "bumps" (Bakker, 2004), that is, central slices of a distribution containing a large proportion of the data, though these categorization schemes rely on natural breaks in the shape of a data distribution rather than equal width "bins".

When people use the "binning" features of software, they typically describe what they're seeing either by general comments on the shape of data, by comparing the number of data points in different bins, or by comparing the percentage of data points in different bins. The multiplicative reasoning, including proportional reasoning, needed to use the percentage strategy is important to thinking well about data, and it has been highlighted by several researchers. Shaughnessy (1992) in his review article claims that the ratio concept is crucial in statistics and often lacking among students (p. 479). Upon re-examining their comprehensive framework for middle school students' statistical thinking (Mooney, 2002), Mooney and colleagues (Mooney, Hofbauer, Langrall, & Johnson, 2001) changed just two elements—they modified the category for organizing and reducing data, and added the important missing element, multiplicative reasoning, which they describe as, "reasoning about parts of the data set as proportions of the whole to describe the distribution of data or to compare data sets" (p. 438). Saldanha & Thompson (2002) argue that seeing a sample as a "quasi-proportional small-scale version of a larger population" is an important conceptual move for students in making statistical inferences. In a 7th grade (age 12) teaching experiment, Cobb (1999) proposed, "our goal for the learning of the classroom community was that reasoning about the distribution of data in multiplicative terms would become an established mathematical practice that was beyond justification" (p. 11) and described a fair amount of reasoning by use of "qualitative proportions" in their analysis. In a subsequent 8th grade (age 13) teaching experiment focusing on looking for patterns in bivariate data, Cobb and colleagues (Cobb et al., 2003) took multiplicative reasoning as the starting point. In fact, the multiplicative reasoning needed to normalize data, that is, to make the scale of numbers the same so they can be compared, is a powerful technique used widely throughout statistics. Examples include rescaling variability in standard deviation units when calculating Z-scores, calculating the mean to yield a per case measure of an attribute, or using percentages instead of counts to deal with differences in sample size, as we will see in this study.

While important in statistics (and elsewhere), proportional reasoning can be difficult, especially for students who are attempting to distinguish it from additive reasoning (Harel & Confrey, 1994). Lamon (1994) details some of this complexity that is especially relevant for data analysis, stating that proportional reasoning requires "unitizing", i.e., "the ability to construct a reference unit or unit whole, and then to reinterpret a situation in terms of that unit" (p. 93), as well as "norming" which includes the idea of percentages, i.e., "reconceptualizing a system in relation to some fixed unit or standard" (p. 94). These transformations require shifting attention from units to relationships among units, a more abstract idea. At the same time, working with these relationships reduces the amount of data to which one must attend at the same time which, Lamon argues (citing Case (1978; 1980)) may "facilitate reasoning [by] easing the load on the working memory" (p. 112).

This paper describes our experiences studying teachers' and students' uses of binning and proportional reasoning strategies using a computer tool that makes each of these strategies more accessible and flexible. Since such tools are new and not yet widely disseminated, we know very little about how teachers' and students' strategies develop when they have these resources available. Which components of the software do teachers use and how do they take advantage of the interactive possibilities afforded by the software? What can we learn about teachers' and students' statistical reasoning by analyzing their interaction with these new tools? What implications can we draw from these data for teaching and learning?

3. CONTEXTS AND METHODS

The data for this paper come from several sources, all connected with the Visualizing Statistical Relationships (VISOR) project at TERC in Cambridge, Massachusetts, USA. VISOR is a teacher professional development and research project studying how people learn about data analysis and statistics and how computer visualization tools can enhance that learning. In VISOR, the professional development and research goals were often mixed. We offered opportunities for teachers to explore data topics such as ways of describing data, stability of measures and the role of sample size, making inferences about group comparison and co-variation situations, and confidence intervals, among others. However, we focused less on *teaching* teachers specific things than on exploring their thinking in the context of use of computer software tools. Teachers explored a variety of data sets using two innovative software tools, TinkerPlots™ (Konold & Miller, 2004) and Fathom™ Dynamic Statistics™ (Key Curriculum Press, 2000). In the group, they also talked about teaching about data analysis, and brought in examples from their own classrooms of their students' thinking and work using these tools. By focusing on how people think about and explore data, the project hoped to help teachers develop a sense of themselves as data analysts, to understand better some of the issues that arise in learning about data and statistics, and to feel more confident teaching about data in richer and deeper ways.

In the VISOR seminar, we worked with a group of 11 middle- and high-school teachers (8 women, 3 men; 6 middle school, 5 high school; 10 White, 1 Black) from mostly urban Boston-area schools, meeting biweekly for three hours after school over the course of two years. In fact, only eight of the teachers continued into the second year. In its third and final year, VISOR worked with a new group of nine teachers. Teachers varied in their comfort with computers and in their prior experience with statistics, some had had very little exposure, a few taught AP Statistics (Advanced Placement high-school courses that provide college credit) or had done data analysis in industry. While some taught semester- or year-long statistics courses, most only taught about data and statistics during a few weeks each year.

We videotaped group sessions, took extensive field notes, and collected teachers' work from the seminar, including video feeds from the computer work of one small group each session. After each session, we created a rough log of the videotape, developing more detailed transcripts for certain sections as needed in our analytic work. We also observed teachers in their classrooms, and collected field notes and examples of students' work, as well as copies of what teachers brought in from their classrooms. Several times during the two years, we conducted formal, individual, audio- or videotaped interviews with teachers on a variety of topics. Finally, one of us conducted a 13-week

teaching experiment on data analysis with a group of 12 relatively advanced sixth grade students (age 11–12) from an urbanized suburb of Boston, taking field notes and reflective teaching notes after each session, and also collecting examples of student work.

Our research goals and methods were primarily descriptive and exploratory, within the goal of discovering how teachers used new capabilities of TinkerPlots™ to compare groups. The authors, who collaboratively led the seminar as well as the research, met regularly to discuss teachers' prior work, to puzzle through what the data showed about how different teachers were making sense of the problems, and to plan sessions to illuminate and highlight different conceptions. These analyses most resembled a combination of group clinical interview (Clement, 2000) and teaching experiment methodologies (Steffe & Thompson, 2000). The authors were sometimes joined in these discussions by Bill Finzer (designer of Fathom™) and Cliff Konold (designer of TinkerPlots™) to focus on aspects of the software tools that might affect teachers' thinking. We also met regularly with a research team at TERC to more closely analyze the formal, transcribed interview data. In this process, pairs of researchers separately analyzed each transcript using a combination of etic codes developed from our theoretical frameworks, and emic codes that emerged from the interviews themselves (Miles & Huberman, 1984; Patton, 1990). We wrote memos about each participant and discussed discrepancies in our analyses until we reached agreement. We then compared across participants to look for common themes and methods, as well as for interesting variations.

3.1. SOFTWARE TOOLS

While we used both TinkerPlots™ and Fathom™ in VISOR sessions, the TinkerPlots™ software provided the platform for the examples we will use in this paper. TinkerPlots™ is a data analysis environment primarily for middle-school classes that provides students with a wide range of tools to create traditional and non-traditional data representations. By various combinations of sorting and separating data into categories, ordering or highlighting information by the value of an attribute, and stacking and otherwise organizing data, users can make graphs that are both familiar and very different from those typically seen in school or, for that matter, in most research settings. Users can display numerical information about plots: the value and position of the mean, median, mode, or midrange; the number or percentage of cases in bins or between sets of moveable dividers; or the value at a moveable horizontal or vertical line. The software offers tools for displaying data as value bars (in which the length of each bar represents the magnitude of the value of an attribute for a case), or fused into rectangular or circular areas. Finally, it provides tools for creating box plots, as well as innovative "hat plots" that partition the data like box plots, but based on user specifications such as percentages of the range or of the data, or numbers of standard or average deviation units from a center.

3.2. DATA SETS

While our work with teachers and students involved exploring a wide variety of data sets, the results we present in this paper focus (primarily) around two data sets. The first, explored in the middle of the first year, was invented but realistic data, modified from Cobb et al. (1999), comparing the efficacy of two drug protocols for treating patients with HIV-AIDS. Of the 232 patients in the sample (160 men and 72 women), 46 randomly received an Experimental protocol and 186 received the Standard protocol. Outcomes were measured in patients' T-cell blood counts and teachers were given information from an AIDS education group stating that normal counts ranged from 500 to 1600 cells per milliliter. We added a gender attribute to the original data designed in such a way as to show an interaction between gender and protocol in their effects on T-cell counts, that is, differences across categories and an interaction of categories in a numerical attribute.

The second data set, explored early in the second year, consisted of real survey data of 82 students (34 girls and 48 boys) attending two western Massachusetts high schools (51 from Holyoke and 31 from Amherst) collected by Cliff Konold in 1990 and included with TinkerPlots™ (US Students: Konold & Miller, 2004). Attributes include students' height and weight, number of older and younger

siblings, hours spent doing homework and working, and average grades received, among others. Teachers focused on school and gender differences in grades received and on hours per week spent doing homework, that is, differences across categories in a categorical and a numerical attribute.

We will also look briefly in this paper at two data sets exploring the relationship between two numerical attributes, both of which were explored late in the second year of VISOR. The first compares the median age in each state with the percent of the state population voting for George W. Bush in the 2000 U.S. Presidential election. The second looks at the relationship between average state educational spending and number of teen births per 100,000 population in the U.S.

These data sets offering different types of data and relationships among data allowed us to see a range of ways that teachers and students thought about and dealt with the variability in data using TinkerPlots™.

4. RESULTS AND DISCUSSION

We discuss in this section results pertaining to our research questions and the relationships between them. We describe how teachers with diverse statistical analysis and teaching experiences approach issues of variability, especially when comparing groups with one numerical and one categorical variable. We also document in less detail examples of looking at the relationships between two numerical variables. Throughout this section, we also relate these findings to our second research question, regarding the affordances offered by a statistical visualization tool like TinkerPlots™.

The purpose of describing *teachers'* use of TinkerPlots™ in comparing groups is two-fold: First, it suggests the kinds of strategies that *students* might use as well when they have TinkerPlots™ as a resource. In fact, as described below, we have confirming evidence from observations in classrooms that students and teachers share these approaches. Second, it helps us learn how teachers approach comparing groups tasks and, therefore, what they need to learn to guide a statistically meaningful conversation for their students.

The teachers in the VISOR seminar created many previously unseen (at least by us) graphs and were extremely creative in their approaches to comparing groups in the data sets described above. In general, our results confirmed Konold et al.'s (1997) observation that students (in this case teachers) seldom use a measure of center as their first method for comparing two data sets presented in graphical form. We report here on two key types of strategies that teachers used in comparing groups with TinkerPlots™, describe how essential design features of TinkerPlots™ influenced these strategies, and comment on whether and how these techniques developed as the VISOR seminar went on. In addition, we describe evidence that students' use of TinkerPlots™ brings up many of the same data analysis issues that arose in the teacher seminar. We situate these descriptions in the general dilemma teachers and students are facing: How to make comparisons between two groups when each of them exhibits considerable variability.

The primary strategy we analyze is one that is both new and unusually easy to use in TinkerPlots™, analyzing a distribution by dividing and chunking it into several pieces. While many statistical analysis packages provide ways to create histograms and even to change their bin sizes in a limited way, it is the flexibility that TinkerPlots™ provides for manipulating and observing information about sections of a distribution that creates a new and powerful tool for learners. Dividing a distribution into bins is effortless in TinkerPlots™, as this representation is a primitive among the graph operations TinkerPlots™ makes available. The strategy of creating multiple "bins" along an axis effectively reduces the number of actual values in the distribution, thus reducing apparent variability.

The need for the second strategy we describe arises in some part from using the binning strategy in the context of comparing groups. If any two groups are of unequal sizes, the binning process will require an understanding of proportional reasoning to compare the two groups. In this context, we see both teachers and students struggling with the difference between additive and multiplicative reasoning. The first pays attention to the *number* of points in a bin, while the second focuses on the *proportion* of points in each bin. TinkerPlots™ supports both of these kinds of reasoning in slightly different ways. We analyze here the ways in which TinkerPlots™' features affect students' and

teachers' uses of additive and multiplicative reasoning and explore the new strategies and dilemmas that arise from the use of TinkerPlots™ features.

4.1. ANALYZING DATA IN BINS

One of the most common techniques both teachers and students used to compare groups consisted of limiting the number of unique values in the data sets by creating two or more bins. Considering a set of data points as a relatively small number of groups of values is not a new idea; in fact, it is the basis for many statistical representations. Histograms and box plots both partition data sets into groups within which all values are essentially the same for analytical purposes.

The role of bins in TinkerPlots™ is notable because the software automatically goes through a "binned" stage as data are separated according to the value of a highlighted attribute; the user does not need to imagine or specifically request a graph with bins. Thus, many more of the graphs produced in our seminar used bins than might have been the case with other software tools. This representational strategy is not specific to TinkerPlots™. In fact, we have seen similar binning approaches when teachers and students analyze data with pencil and paper. But the immediacy of TinkerPlots™' binning functions affords more complex strategies. We describe below some of the common ways teachers and students used bins to compare groups.

Each of the methods described below highlights some of the consequences of regarding data in bins. They are all examples of a tension inherent in data analysis; finding the right balance between the advantages of reducing variability as a way to deal with the complexity of data, on the one hand, and the risks of making claims that might not be true, or would have to be qualified, were all the data included in the analysis, on the other. Several of the examples below also illustrate the interplay of contextual and strictly numerical ways of looking at data, which we frequently observed in teachers' discussions of their analyses.

Using cut points, both system- and user-generated

One of the simplest ways to create bins in a distribution is to divide it into two parts. We have used the term "cut point" to designate a value in a distribution which divides it into two groups above and below that point. When a user begins to separate the values of a variable, TinkerPlots™ immediately provides a single cut point using the software's rule of using a value roughly at the rounded midrange.

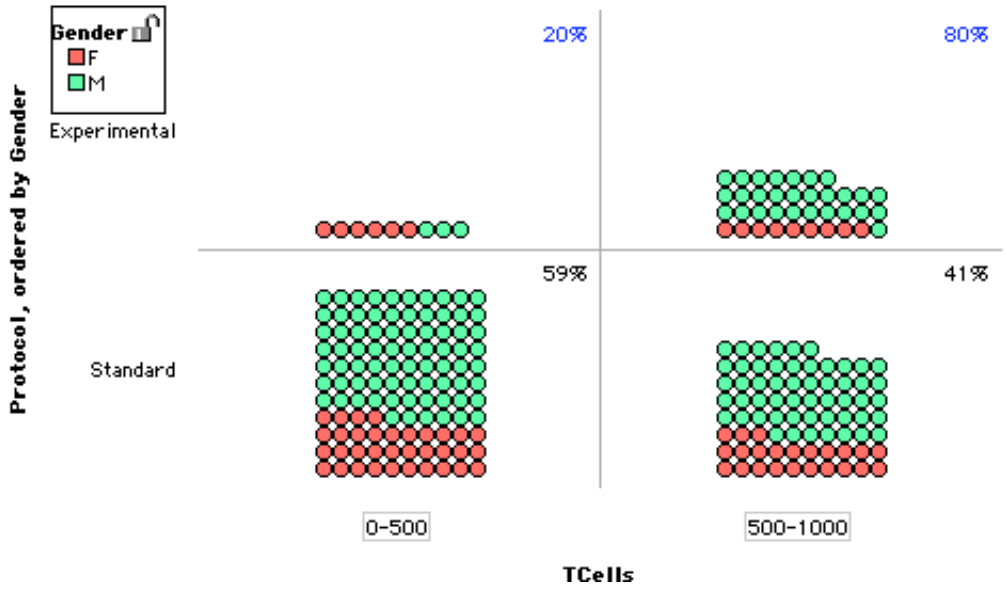

Figure 1. Comparing the percent of each protocol above a single cut point

If the data are split into two subgroups as well, we get a two-by-two representation (see Figure 1). A legitimate comparison between the two groups, then, would involve comparing the percentage of each group above and below the bin boundary. Understanding the importance of using percents rather than counts is an issue we discuss in Section 4.2, below. In fact, TinkerPlots™ supports this comparison by making it possible to display the percents in each bin. In Figure 1, an argument can be made for the superiority of the experimental treatment based on the larger percentage of experimental cases in the high T-cell bin. Remember, high T-cell counts are better.

While this kind of representation was common early in our seminar because it was so easy to create, teachers quickly grew to reject this representation because it hid so many of the details of the shape of the data. A more common representation was Figure 2. Here we see that the teacher created seven bins and displayed the percents in each. In this particular case, the analysis proceeded by adding together the percents in the bins above 500 T-cells per ml for each protocol, replicating the analysis in Figure 1. At other times, however, these kinds of multi-bin graphs were analyzed in very different ways, as described below under "Comparing slices."

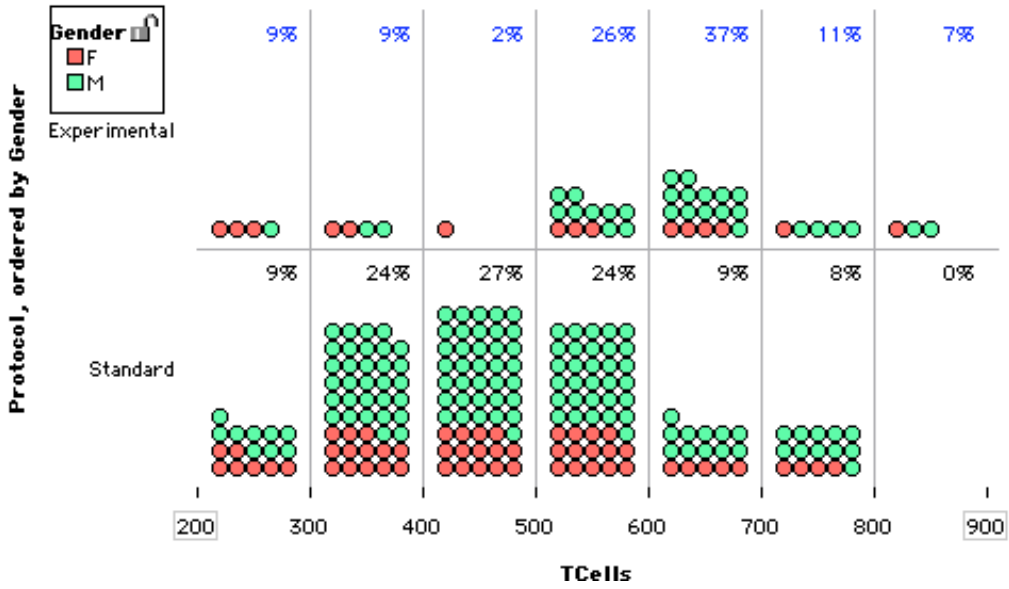

Figure 2. AIDS T-cell data in seven bins with row percents

A more sophisticated approach using TinkerPlots™ was to use a different tool, *dividers*, that allows the user to specify exactly where she wants to put one or several cut points. In Figure 3, the teacher has placed the cut point at exactly 500 using a divider tool to split the distributions into two parts. This representation is different from a binned representation in that the user has to be explicit about creating and moving a divider (by grabbing the small white square at the right side of the page). Note that in this case, the default setting at which TinkerPlots™ drew the first bin and the cut point value that this teacher has chosen are the same: the difference is that the teacher has detailed control of the dividers and could change the cut point to 495 or even 497.

Representations with dividers arose fairly early in the seminar, although it remained difficult for some teachers who had trouble mastering the tool and therefore continued to use primarily bins in their analyses. This representation is interesting because, while imposing a cut point, it also retains visual information about the rest of the shape of the distributions. Unlike Figures 1 and 2, the exact T-cell value of each point is represented on the graph. This became a preferred graph for some teachers who would routinely "separate a variable completely" (create a representation without bins), then impose their own dividers. Figure 3 is also different from the rest of the graphs in this paper because the continuous variable is displayed vertically, but that is not unusual in TinkerPlots™, since it is equally easy to put a variable on either axis.

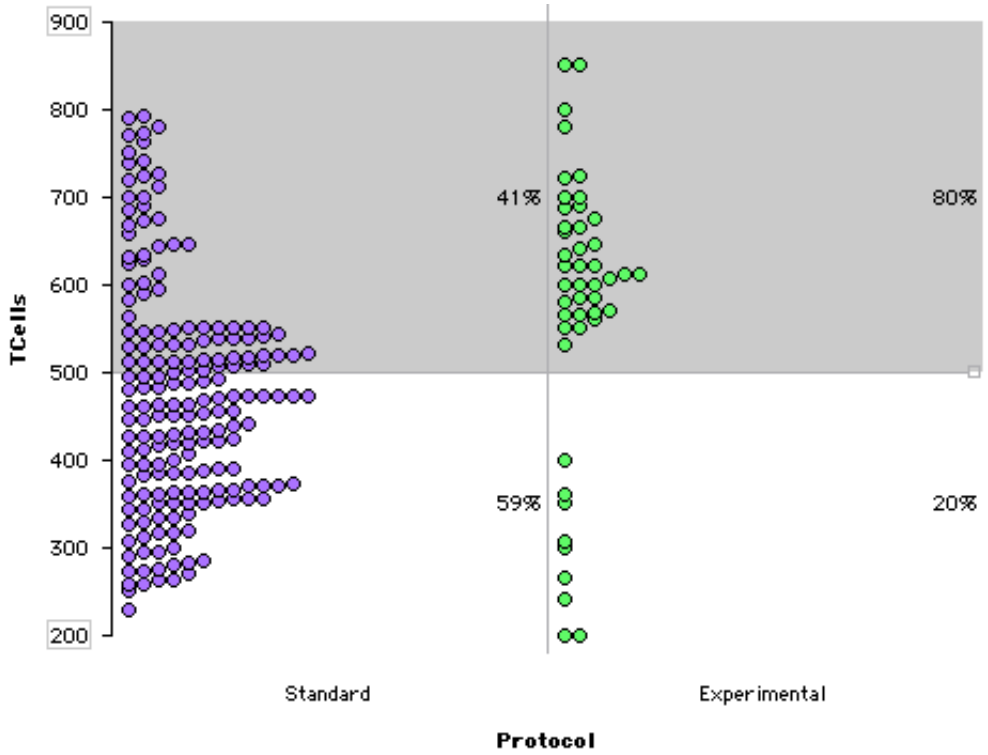

Figure 3. Using a divider to compare percentage of T-cells above and below a user-specified cut point

This description of three graph types is not intended as a developmental sequence, as different teachers made different choices in representations at any point, and an individual teacher might have alternated among several kinds of representations. There are, however, a few general points to be made about trends in teachers' choices.

Most teachers did prefer Figure 2 over Figure 1 (at least, after the first week or two), even though they performed essentially the same analysis with each: adding up the percentage of cases with values over 500. In fact, using Figure 2 to make the argument requires more work than Figure 1, since the percents above 500 are not yet totaled. Our observations lead to this hypothesis: The teachers used Figure 2 so that they could CHOOSE the cut point based on the shape of the data, then use their chosen cut point to compare percents across groups with regards to that cut point. Working with a cut point from Figure 1 runs the risk of choosing a value that would seem more or less representative of the data depending on the distributional shape. Figure 2 combines the ease of using bins with a desire to see the shape of the data. In this case, we hypothesize that if the distribution were less symmetrical, teachers may have wanted to use a different approach than the single cut point one. Some tentative findings from research not reported here support this hypothesis.

What influenced a teacher to choose between Figures 2 and 3 in analyzing a data set? As described above, teachers did not move from using only a graph like Figure 2 to using only a graph like Figure 3. However, it often happened that when teachers "completely separated" a graph, such as that in Figure 3, they then re-binned the data (which one can do easily in TinkerPlots™). Our hypothesis about this tendency is that, faced with a large number of points stacked unsteadily on the axis, teachers often chose to retreat to a representation with more built-in structure and less apparent variability. Interestingly, in none of these examples were teachers likely to look at a measure of center such as mean or median.

All of these graphs were generated as part of the teachers' discussion of the AIDS data, which took place after three months of the seminar. In generating and making arguments from graphs such as Figures 1 through 3, teachers often used the value 500 as a cut point, since that was the lower end of the normal range of T-cell counts according to the information sheet we had given them. Using

these representations, teachers argued in support of the Experimental treatment's superiority by noting that the percentage of patients above 500 was greater in the Experimental group than in the Control group. Using graphs such as Figure 3, teachers noted that, "The Experimental treatment yields a higher percentage of participants in the normal range (above 500). [Figure 3] shows that in the Experimental protocol, 80% of participants were in the normal range, while in the Standard protocol, only 41% of participants had T-cell counts above 500." One teacher was more effusive, saying, "A huge preponderance don't get much better than 500 in the Standard, though a huge preponderance do better than 500 in the Experimental."

Exploring these three classes of representations led to a discussion in the group about the rigidity and arbitrariness of the cut point, a discussion that included both contextual and statistical arguments. Some teachers wondered how rigid the value of 500 was from a medical perspective, would T-cell counts of 495 mean someone was *nearly* healthy? What about 450? One teacher made an analogy with grouping people by height saying, "If somebody was right at the borderline that doesn't mean they're short." Other teachers speculated that there *could* be a biological mechanism that created a sharp distinction in sickness rates once T-cell counts reached a specified level. "There may be a magic number that's very close. To get infected by a germ, there has to be a minimum number. If you get less than the minimum number, you get nothing." Teachers also made arguments for the significance of the cut point 500 grounded in the shape of the distribution. Some people noticed that there was a large gap in the data for the Experimental group just below 500, which gave them confidence in using this value to divide the data. This, in turn, led to further context-driven speculation about a possible genetic difference in response for the people at "the low end of the Experimental protocol." The back and forth between patterns observed in the data and those driven by the context led the participants beyond what they would have seen by just looking at the distribution as numbers. After they had all seen the entire distribution, most teachers seemed comfortable with reducing variability by imposing a single cut point, although there was some disagreement about exactly where to put it, thus focusing on this one distinction rather than trying to integrate and deal with the variability of the entire data set at once.

In some ways, all of these uses of cutpoints can be seen as examples of viewing data as a classifier using Konold et al.'s (2003) schema, i.e., attending only to the frequency of data in a particular bounded range, e.g. below 500. From another perspective, though, comparing percentages of the two groups above or below a particular value can be seen as paying attention to aggregate features of *all* of the data since, paraphrasing one student speaking about the Amherst-Holyoke data set, 'knowing that 55%…are above the cut point means that 45% are *below* the cut point.' A comparison of *counts* above a cut point across two distributions, however, does not take all of the data into account except in the rare case of equal sized groups—an important distinction that we will discuss in more detail below.

There are several ways that cut point reasoning can go awry, however. When cut points distinguish only a small portion of a data set, just a few points on one side of the cut point and the rest on the other, they essentially serve to identify unusual values or outliers and conclusions based on these subgroups may not be robust. In addition, cut points only describe an aggregate view of data when comparing the percentage of data on either side of a *single* cut point. We contrast this use of a single cut point with a form of pair-wise comparison of values between two groups that uses "slices" of a distribution, i.e. portions of a distribution formed by *more than one* cut point, so that the distribution is divided into more than two sections. Performing a comparison using several slices is, indeed, using a classifier view of data. The difference between "slices" and the divisions formed by a single cut point is that a single slice in the middle of a distribution effectively disregards everything else about the distribution—information that could radically change an interpretation based solely on the data in the slice itself. Several examples of these problems with slices follow.

Comparing slices

Although we have no examples of teachers exploring the AIDS data using pair-wise comparisons across internal slices, we have seen both teachers and students make such comparisons using the

Amherst/Holyoke data set. These observations mirror those that other researchers (Konold & Pollatsek, 2002; Watson & Moritz, 1999) have noticed in students comparing groups.

For example, when two teachers were using the Amherst/ Holyoke data to answer the question "In which town do students spend more hours doing homework per week?" they proceeded to divide the data into seven bins. They then agreed to discount both the bottom bin (less than four hours) and the top four bins (greater than 12 hours) to focus just on the two "middle" bins in which most of the data were clustered (see Figure 4). They argued that they had "clear explanations" for what was going on with both the bottom and top groups, so that it was acceptable to exclude them from the analysis. Specifically, they argued that the number of students doing fewer than four hours of homework a week was the same in both schools and could be discounted because it would add nothing to the comparison. By contrast, they said, "The top is a lifestyle no matter where they live," and therefore should also be discounted because it didn't represent something about the typical student. Here is another example of teachers' preferring binned data (as in Figure 2) to fully separated data (as in Figure 3) so that they could understand the data in easily visible chunks, even though they often just accepted the chunks that the software provided.

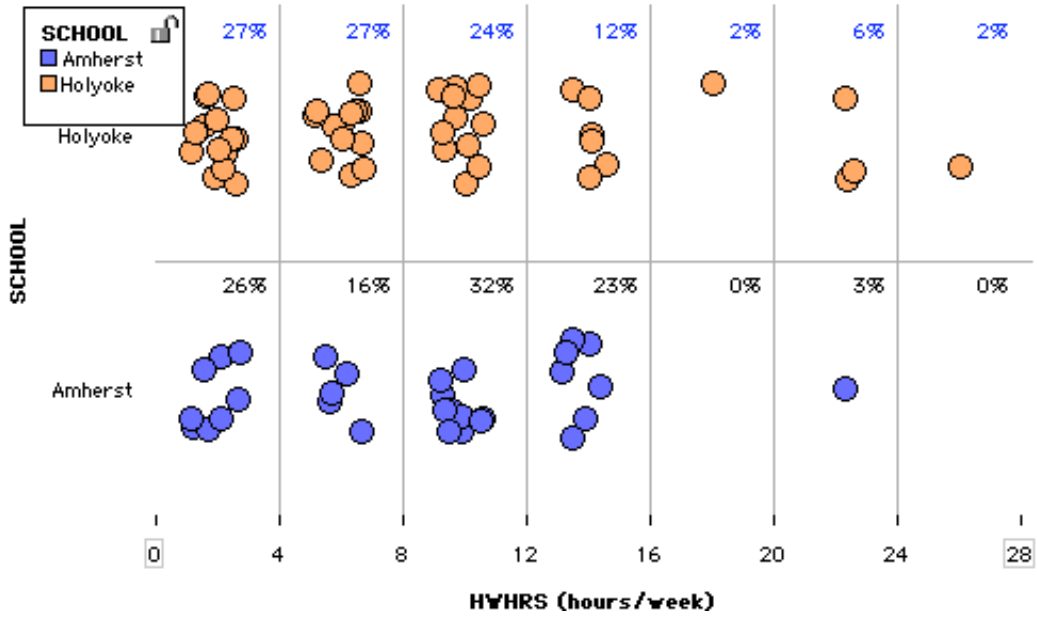

Figure 4. Looking at a slice of Homework hours in two Schools. Teachers focused on students who did between four and twelve hours of homework each week.

Having constructed this graph and chosen to focus on just the "typical" students, the teachers compared the percentage of students from each school in each of the two bins they found interesting. In the 4- to 8-hour bin, they found 27% of Holyoke students and just 16% of Amherst students; in the 8- to 12-hour bin, they found 24% of Holyoke students and 32% of Amherst students. These two comparisons within slices led them to conclude that Amherst students did more homework than Holyoke students. In fact, they were comparing the comparisons, noticing that a higher percentage of Holyoke students were in the 4- to 8-hour bin, and a higher percentage of Amherst students were in the higher, 8- to 12-hour bin. They reasoned that, since Amherst had a higher percentage of students who studied eight to 12 hours and Holyoke had a higher percentage studying between four and eight hours, Amherst students must do more homework. The teachers also looked at the data another way, noticing that the ratios of students in the two bins within each of the schools was different—roughly equal numbers in Holyoke study "8 to 12" and "4 to 8" hours respectively, whereas twice as many Amherst students study "8 to 12" hours as study "4 to 8" hours. This confirmed their view that Amherst students studied more than Holyoke students

However, there are problems with their argument. A slice-wise comparison across groups when there are multiple slices effectively ignores the rest of the distribution, i.e., if you only know about the

percentage of students in each school who study between four and eight hours, you can't really say much about the overall pattern. As stated, the conclusion doesn't account for differences in study habits of those who study more than 12 hours who, if included, might lead to a different analysis and conclusion.

We can view the teachers' approach to this analysis from the perspective of reducing variability. By putting the data into bins, they reduced the overall variability to just seven categories. By using both data-based and contextual arguments to discount the relevance of several of these bins, they further reduced the variability in the data and, in the end, compared just four numbers. This level of detail was sufficient for them until they were asked *how much more* Amherst students studied than Holyoke students. They then found their representation and analysis insufficient to answer the question. In fact, in trying to come up with some kind of number, these teachers wondered whether the horizontal position of points in each bin had any quantitative meaning, and had to be reminded that bins served merely as categories without distinguishing among their members.

Dividing distributions into multiple bins, then making an argument to disregard those on the extremes, was a common approach throughout the seminar, both using TinkerPlots™ and on paper. It is, we conjecture, an extension of the strategy of "disregarding outliers." In fact, one of the pieces of "statistical knowledge" with which several of the teachers entered our seminar was: "If you're using the mean, it's important to disregard the outliers." Or, differently stated, "If there are outliers, use the median." With a TinkerPlots™ binned representation, it is simple to disregard entire groups of "outliers" because there is no visible difference in value among points in the same bin.

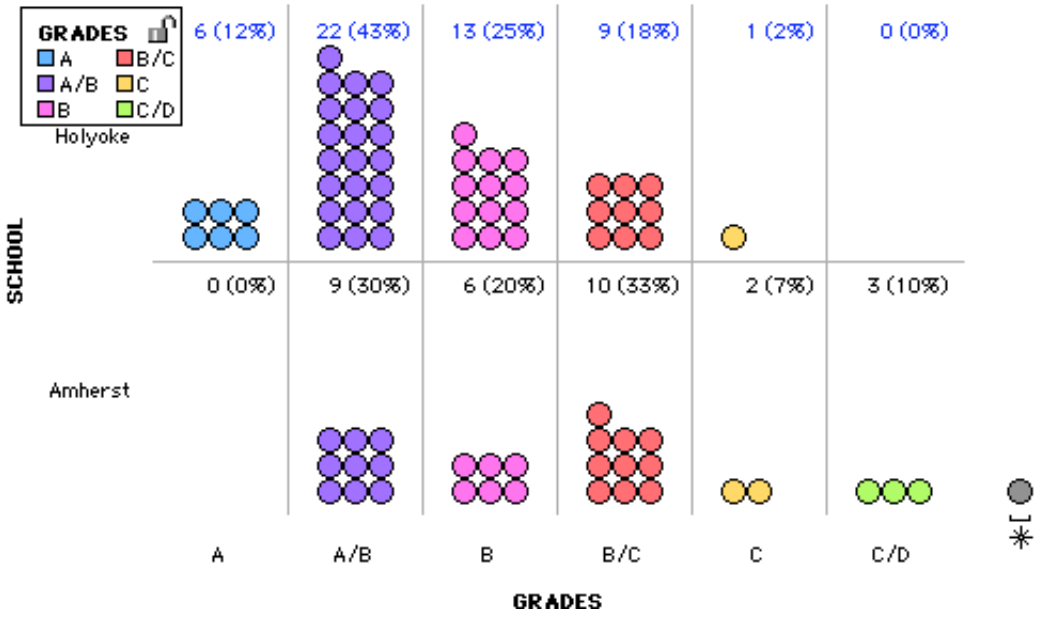

Figure 5. Comparing Grades in Schools by looking at the A/B slice

Another example of people looking at slices to compare distributions comes from the 7th grade class of one of our participating teachers. Using the same data set, but looking at grades received by students across the two schools, one student looked only at the students who received a mix of As and Bs (the A/B column, second from the left in Figure 5) to argue that more Holyoke students got better grades than Amherst students, although we don't know if she was referring to numbers or percents as they're both displayed and both point in the same direction. Likely, this student was focusing on the highest category about which she could make a comparison, since no Amherst students got straight As, though such a focus on a single slice ignores the variability and shape of the distribution.

In contrast to focusing on an internal slice created by more than two cut points, some students used a *single* cut point to split the data set between the A/B and B categories so they could compare the percentage of students in each school getting either As or As and Bs (55% Holyoke, 30% Amherst). They argued persuasively from that observation that Holyoke students got better grades. However, others thought such a comparison wasn't fair, that "you shouldn't pay attention just to the

smart kids," and wanted to look at the students getting Bs or Bs and Cs. They seemed to be torn between a desire to look at *more* of the data, expanding their view to include kids more in the middle *in addition* to the "smart kids," and wanting to *limit* their view by looking at a single slice, such as just those students getting Bs.

Another student in this discussion proposed expanding the number of students being considered by moving the cut point to just below the B range. That would mean that 80% of Holyoke students would be included and 50% of Amherst students, which seemed both like a large enough fraction of the students in the sample, and a large enough percentage difference to be able to draw the conclusion that Holyoke students get better grades than Amherst students. One could argue, although these students didn't, that setting the cut point below B is really more of a comparison of the lower end of the distribution than of the upper end, i.e., which students get worse grades? Again, we see a tension in students' techniques between, on the one hand, *narrowing* their perspective on the data by using bins to reduce the variability and number of points they have to attend to and, on the other hand, *expanding* the scope of data they're considering to include a minimum number of students with which they feel comfortable.

The ease of creating bins in TinkerPlots™ supported both cut point representations and the possibility of focusing just on internal bins as "slices," which produced a kind of argument that we conjecture would not have occurred as often otherwise. For some middle school students, the distinction between slices and cut points remained problematic. In general, the distinction between slices and cut points presented fewer problems to the teachers in our seminar after the beginning, but one teacher continued to routinely disregard the ends of a distribution for contextual reasons and focus on the middle. For example, in analyzing the weight of student backpacks before and after a hypothetical assembly on the topic of back problems caused by heavy backpacks, it was this teacher who disregarded those carrying less than 4 pounds as being "slackards."

4.2. PERCENTAGES AND PROPORTIONAL REASONING

Working with unequal size groups brings up the issue of additive versus multiplicative reasoning. Figure 6 illustrates the difference between additive and multiplicative reasoning. In this binned representation, both numbers and percents are displayed in each bin. Thus, it is possible, and even made relatively simple by TinkerPlots™, to compare the *number* of subjects with T-cell counts above 500 in the Standard vs. Experimental subgroups (using additive reasoning). In the case of Figure 6, this would lead to the incorrect conclusion that the Standard protocol was more effective because there are more subjects above 500 in the Standard condition. Of course, the correct way to make this comparison is by using percents, using multiplicative reasoning. Several of the teachers in our seminar correctly compared the two drug protocols by looking at the *percentage* of patients with T-cell counts above the "healthy" cutoff of 500. Interestingly, however, several teachers in our seminar struggled with the distinction between an analysis based on numbers of points and one based on relative percents. The tendency to use counts rather than ratios was surprisingly robust.

Another task we gave teachers was to judge if the Experimental treatment was as good for women as it was for men. Figure 6 is one representation that could support this analysis. In each bin, every man is colored green and every woman is colored red. By visually separating the males and females in each bin, teachers looked at the rough proportion of men to women in each of several ranges of T-cell counts. A further step one group of teachers took was to create a circular fuse in each bin to create pie graphs (Figure 7). These two representations, though, provide very different views of the data, since in Figure 7, the salient part of the representation is the ratio of men to women (green to red) and the number of cases in each bin is displayed only as a number. In Figure 6, however, we *see* which bins have more data, but the pattern of proportions is less apparent. Note that Figures 6 and 7 have different bin boundaries than Figures 2 and 3; depending on the sequence of steps the user has taken to arrive at the graph, bin boundaries may be placed slightly differently. The user can also specify the number of bins and the bin boundaries.

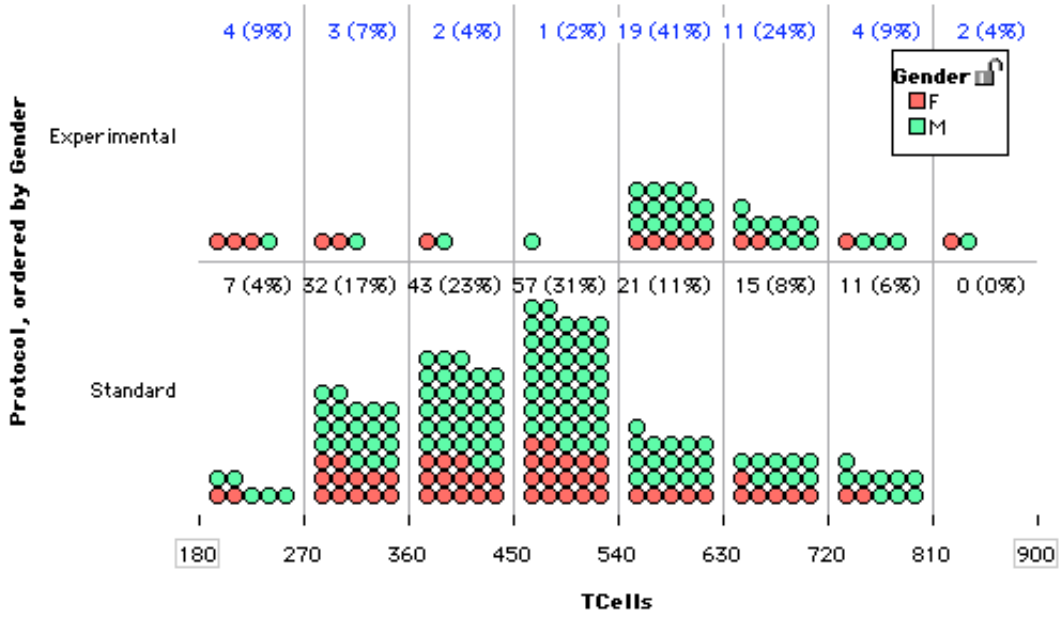

Figure 6. Binned representation showing proportions and numbers simultaneously

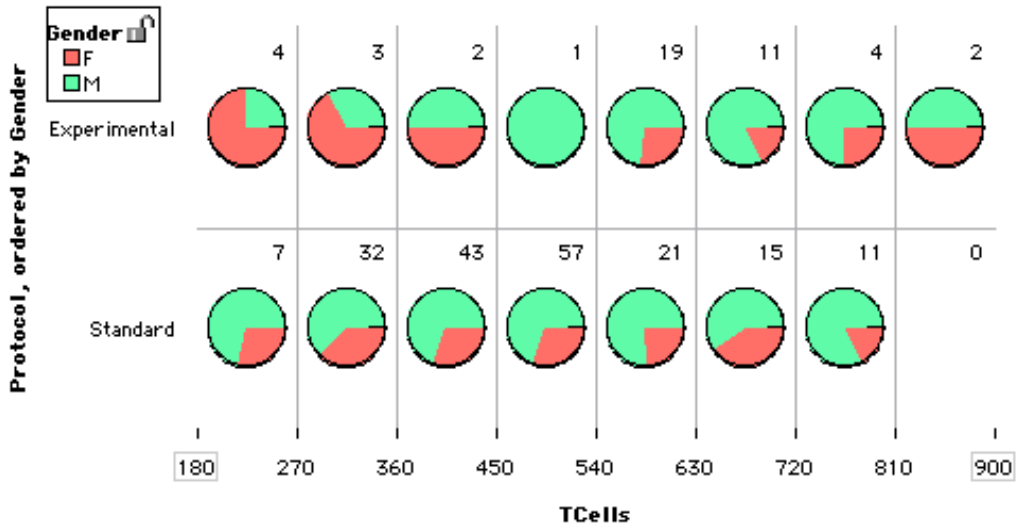

Figure 7. Pie graphs showing shift in gender proportions for Experimental protocol and rough consistency for Standard protocol

Teachers made an unusual argument using Figure 7. Some teachers used these pie graphs to notice patterns in proportions across the entire data set, specifically the rough shift of the gender proportions from low to high T-cell counts for the Experimental protocol (more females had low T-cell counts; more males had high T-cell counts) compared with the rough consistency of the gender distribution for the Standard protocol. The pie graphs enabled teachers to emphasize the fraction of the data of each gender in each bin precisely because the numbers were visually normalized into a single circular whole for each bin, even though the numbers in each bin were different. For several teachers, this was a most compelling graph because of the salience of the visual pattern. One teacher said, "I like the circle because it captures the whole….I can see the three-fourths for that region much better on a pie chart than with a histogram."

Even though the counts are displayed in each bin, the pie graph representation makes it difficult to discern information about numbers in each bin, numbers that must be big enough to draw conclusions with any confidence. For example, there is only one person in the 450 – 540 range for the Experimental protocol, so the fact that the circle is all green gets more visual "weight" than it should.

One teacher found this problem quite disconcerting. She said, "You may like it, but I don't. I think it distorts the numbers." Instead, she preferred Figure 6, which she could visually inspect for proportionality, while also being able to see the relative numbers going into that proportion. "I prefer the histogram because I can actually see the counts. With pie charts you think percent[age]s, you don't think numbers." Still, Figure 7 supports a kind of argument that is not salient in Figure 6 and it is an argument that bears consideration.

While this particular graph (Figure 7) was compelling for a subset of the teachers, pie charts seldom appeared after this data set. It is possible that the teachers who didn't like the pie charts because they "distorted the numbers" convinced the others that, despite its simplicity, this graph was likely to be misleading.

As noted above, some learners are troubled by how these representations hide absolute numbers, while others are not. We've seen both students and adults look at sets of pie graphs and forget that some of them represent only three points while others represent 30. When this is pointed out to them, they *do* know that it would be easy to dramatically change the proportions in the set of three by changing only one data point, and therefore they're less likely to trust that proportion than the proportion in a pie graph representing more data points. Still, we have seen many teachers forget about this concern unless it comes up in conversation. Although we do not have enough evidence to know why, we hypothesize two contributing factors. First, several teachers described how easily they saw patterns in a sequence of pie graphs; visually these patterns are much more striking than the numbers in the bins. Second, we also know that at the time the group worked on the AIDS data, their appreciation of the effects of sample size was relatively weak, so they may not have focused on that aspect of the representation.

Using proportions in the form of percentages or pie graphs to equalize groups of unequal size is a powerful and sometimes new idea for students in middle school. In the 6th grade teaching experiment conducted by one of the authors, several students were excited when they realized that by using percentages to compare groups of different sizes, they could "make it seem like they're even." This was much preferable to other ideas they had been considering to deal with unequal group sizes, primarily removing points from the larger group until it had the same number as the smaller (Watson & Moritz, 1999). Students were uncomfortable with this solution mostly because they couldn't figure out which points to cut without introducing a bias.

An interesting related issue arose in the 6th grade group among some students who weren't wholly comfortable with proportional reasoning. When students were using percentages even though they knew that the groups were different sizes, some worried that each "percent" *meant a different thing* in each of these groups—that is, ten percentage points may have been six students in one group, and eight students in the other group. How could they compare percentages when they meant different numbers of students? These hardly seemed equivalent. A similar issue arose in a discussion with a VISOR teacher discussing the money earned each week by a sample of Australian high school students. "It's very confusing because if you're realizing four girls equals seven percent whereas only one boy equals four percent…I mean if I didn't put the numbers, I could have just said, 'Okay, percentage-wise the boys make more [money per week].' But if you look really at how many kids each of those really affect..." In all these examples, we see a tension for both students and teachers between recognizing the power in being able to "make groups even" by putting everything on a common scale of 100, and distrusting that transformation and being drawn back to worrying about absolute numbers.

In these examples, we see teachers and students using proportional representations to deal with the variability of group size that is often encountered in data. We also see them struggling with how to simultaneously retain information about group size which is often put in the background, if not completely hidden, when emphasizing proportional views. TinkerPlots™ provides ways to represent both proportional and count information, with relatively more emphasis on one or the other in different views. Pie graphs, for example, focus on proportions without considering absolute counts. By making all these combinations of representations possible—counts without percents, or percents without counts in bins—TinkerPlots™ provides a wide choice of representations and possible

arguments. This, in turn, forces students and teachers to confront and discuss the conclusions that can be drawn, legitimately or not, from each representation.

4.3. BINNING IN THE CONTEXT OF COVARIATION

The teacher seminar had gone on for several months before we approached covariation. By that time, the strategies described above—binning of various sorts, using cut points, comparing slices, using both additive and multiplicative reasoning—had been thoroughly explored, but only in the context of comparing groups in which a single numerical distribution is partitioned into bins, and compared across one categorical attribute. But how would binning and the relationship between counts and percents play out in the context of covariation where there are two numerical variables?

Interestingly, teachers often extrapolated their binning methods to work in a 2-dimensional covariation situation, taking advantage of TinkerPlots™' tools to easily create bins on each axis, thereby partitioning the plane into a "matrix" of boxes (similar to Konold, 2002). For example, in an analysis of a data set on states, relating percent who voted for Bush in the 2000 election to median age, some teachers produced a graph of this "matrix" form by creating four bins on the X-axis and five on the Y-axis for a total of 20 boxes (see Figure 8). Each cell in the matrix contains a "rectangular fuse" of all the data points that belong there. Each small square in a fused rectangle of data points represents a single point—a state—and its color represents the Bush vote; darker colors represent larger percents, as illustrated in the legend in the top right. The data points have not been ordered within each rectangle, so the colors create a checkerboard pattern.

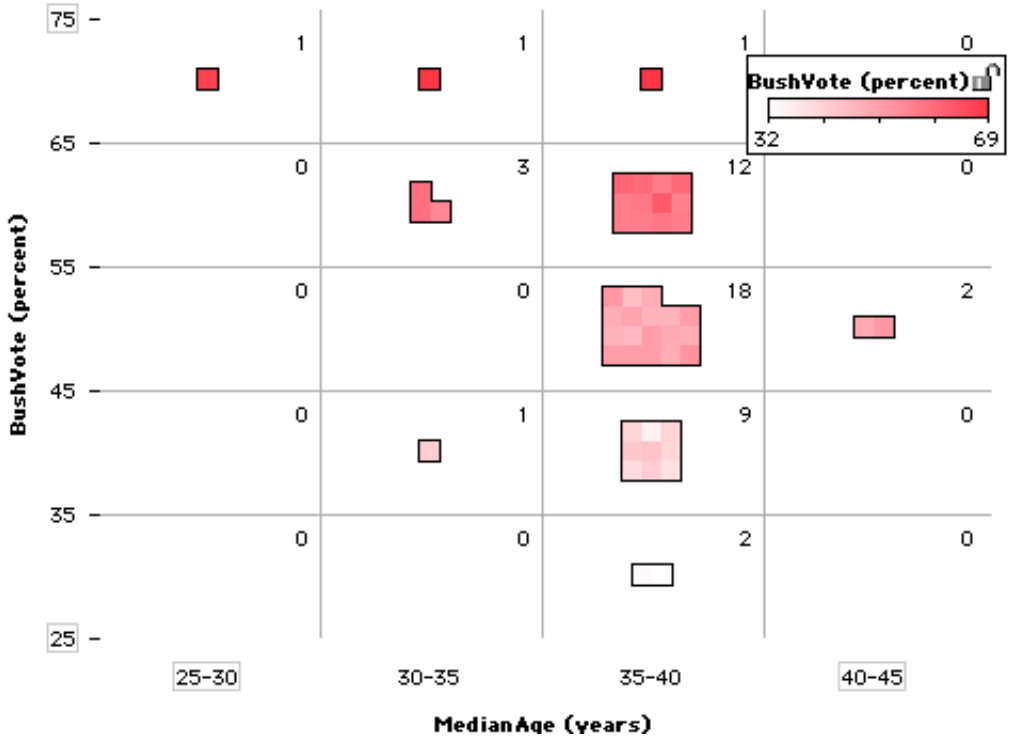

Figure 8. "Matrix" of Median age by Percentage voting for Bush in 2000

Here, the X-axis (median age) is divided into four bins, each representing a 5-year interval. The Y-axis is divided into five bins, each representing a 10% interval of votes for Bush. Thus, for example, the box that includes states whose median age is between 35 and 40 AND which voted between 45% and 55% for Bush has 18 states in it.

Characterizing relationships in these kinds of data is still difficult, even after reducing the variability by binning in this way, and the statements teachers made based on these graphs reflected that. One of the most interesting ways teachers described this kind of graph was to make non-

probabilistic statements of the form, "No country with a life expectancy under 40 has a female literacy rate over 50%." This kind of statement essentially takes the "stochastic" out of statistics and reduces variability still further by finding an "absolute" statement that one might imagine being true for other samples. Using Figure 8, teachers made statements like: "Any state in which the median age is between 40 and 45 has a Bush vote between 45% and 55%," or, "Any state in which the Bush vote is between 25% and 35% has a median age of 35 to 40." In making these statements, teachers were noticing columns or rows of the matrix that have only one cell filled; if more than one cell in a column or row is filled (e.g., the row of states in which 35% to 45% voted for Bush), these kinds of statements can't be made. Note that while these statements are strictly true according to the graph, they are each based on only two states and ignore much of the variability in the data. They are not reflections of the underlying process that may link these two variables, a process that is inevitably noisy and that won't produce exactly the same results every time (Konold & Pollatsek, 2002).

Teachers often migrated towards descriptions that focused on the deterministic, non-stochastic statements they could make from the data rather than the noisy processes underlying the bigger picture. In fact, we've noticed that some people seem to actively avoid making statements about the likelihood of certain features of data being due to chance, preferring instead to handle variability by finding subsets of the data about which they can make more deterministic claims. People want to say: "If I do X, Y will happen. If that's not always true, I can try again with a more precise understanding of the relationship between X and Y." For some people, such a belief in a microscopic version of causality is preferable to the necessity of confronting an inherently variable process. Coming up with a story that predicts exactly some, even if not all, of a data set removes those items from consideration in a stochastic framework.

We might consider the teachers' focus on a small subset of the data in this kind of deterministic way as an example of a "classifier" view of the data (Konold et al., 2003) since the teachers appear to be attending to a small set of data points without considering their relationship to the rest of the distribution. While their thinking does have some "classifier" characteristics, these teachers are thinking in a more complex way. They have *some* awareness of the rest of the data set, since the cell being described must be picked out from, and is therefore seen in relation to, the other cells in that row or column. Still, like cut points that isolate only a few unusual points, this kind of a view doesn't consider much of the data at once. Using Figure 8 to create an argument of this kind does not create an overall description of the relationships between the variables.

There are other examples, however, of a covariation "matrix" in which this kind of argument would be more defensible. For example, using the same States data set, one of the teachers produced the graph shown in Figure 9. In describing this graph, this teacher called attention to "the empty quadrant" in the upper right, which enabled him to say something like, "If a state spends at least $6789 per student, its teen birth rate will be less than 46 per 100,000." The teacher created this graph by placing horizontal and vertical lines at the means for each variable and then felt comfortable making the statement even though there are actually three states where educational spending is above $6789 but whose teen birth rate is more than 46 per 100,000. We conjecture that it was the *form* of the argument he was concerned with, more than the exact details; he knew that there were values of Ed_Spending and TeenBirths for which a statement like his would actually be true.

In fact, other teachers who made similar graphs used slightly different lines in order to completely isolate the "empty quadrant" so that it contained NO points. Is the teacher who left in a few points more comfortable with variability than those who excluded all points before drawing conclusions? Note that while this is a deterministic statement similar to the one above relating median age to percentage voting for Bush, it takes into account characteristics of the entire data set, since there is a significant visual "hole" in the graph, a more global than local phenomenon. And, while this teacher did not explicitly describe these data as representing a signal and noise, one can imagine his statement turning into an appropriate stochastic claim.

It is interesting to note that Figure 9 does in two-dimensions what Figure 3 seems to do in one-dimension. It retains a detailed display of the overall shape of the data while marking and focusing on important or interesting regions. That is, both graphs involve a fully separated view of continuous variables and use dividers set at a contextually relevant value (in Figure 3, a T-cell count of 500), or

reference lines set at both mean values, and then online "pencil" marks (Figure 9) to point to or get information about the graph and thus, the data. While comparable information can be obtained using bins, among our teachers, use of pointing tools on top of a fully separated display is more consistently connected to global statements about the data than are binned representations. A display of the full shape of the data seems to lead to a somewhat more holistic, aggregate view; or perhaps it just doesn't *also* support a classifier view. Yet, these types of uses of tools are less common and perhaps more difficult to conceptualize than are methods of binning. It's not clear whether the use of certain tools enables more sophisticated thinking, or whether teachers and students who have more sophisticated ways of thinking use specific tools to express their ideas.

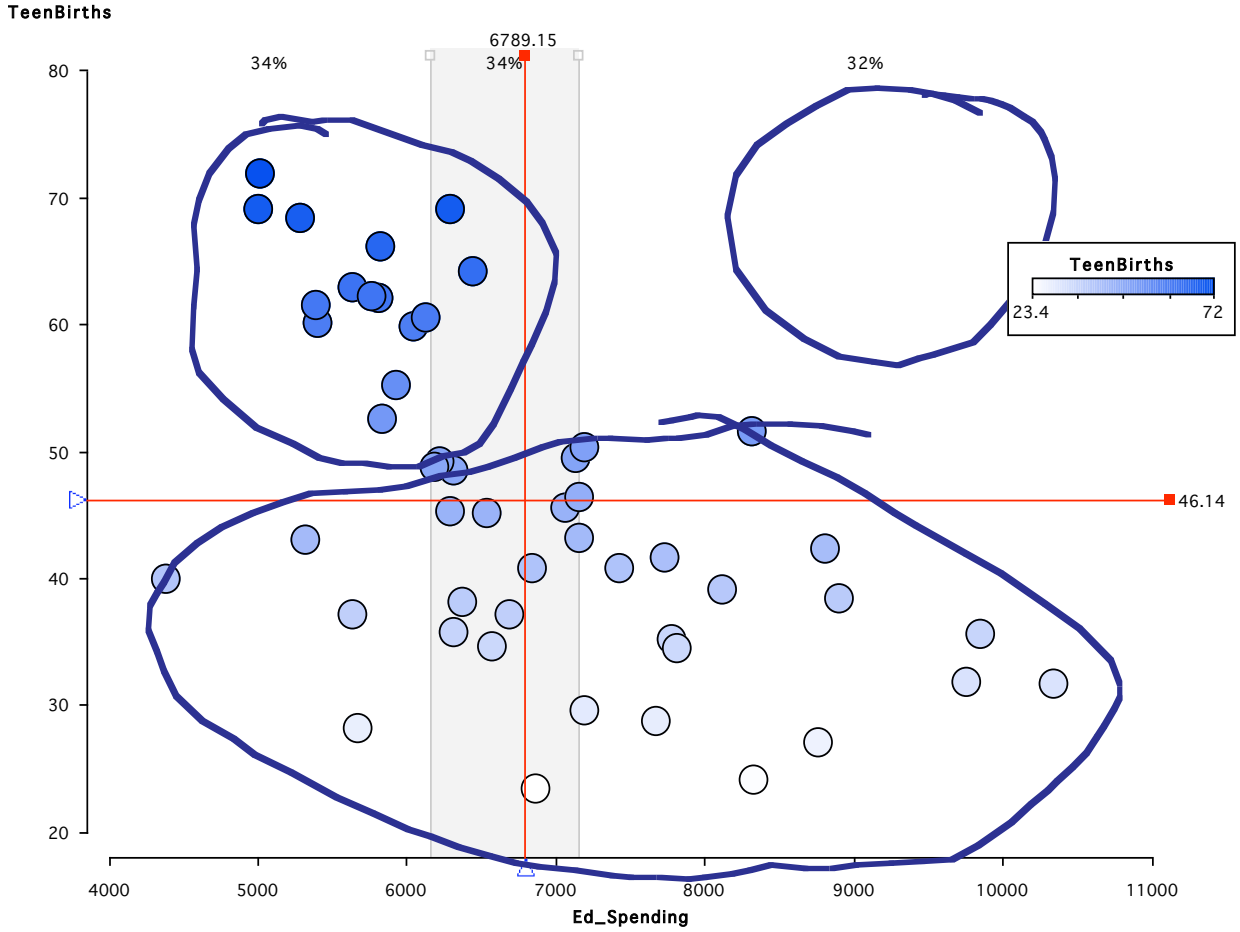

Figure 9. Finding an "empty quadrant" in the data (the circles are the teacher's)

5. CONCLUSIONS AND IMPLICATIONS

This study was designed to examine the strategies that teachers and students use to handle variability in data distributions, and to explore the possibilities for data representation and analysis that new statistical visualization tools make available. Analyzing data with such tools, in this case TinkerPlots™, makes visible reasoning that we could not have observed before. Much as new media both support the emergence of new ways of creating art and may reveal more of the artist's process, these new tools enable new representations and in so doing give us a window into the reasoning behind them. The tools offer the opportunity to examine reasoning strategies that build on the new representations they afford, as well as provide the cauldron within which these strategies emerge.

Our observations of teachers dealing with variability in the context of comparing groups agree with those of Konold et al. (1997): using measures of center was by far less common than the other strategies we have described in this paper. Our experience is, in fact, that *seeing* a distribution makes

it harder to accept a measure of center, especially a mean, as being representative of the entire distribution, with its particular spread and shape. In this sense, the binning capacity of TinkerPlots™, we believe, filled a "need" these teachers had to describe a distribution when the variability was perhaps more compelling than any measure of center. So our focus on binning was both because it was a new and very visible function of TinkerPlots™ and because we saw teachers using it in ways that helped solve a data analysis problem for them. Thus, we saw teachers using binning from early on and as time went on, they continued to use bins but became more sophisticated in their use of them. They became able to indicate more specifically which bin lines would help them make their points or to use the more flexible, but more complex, dividers to make their argument for a difference between groups (e.g. see Figure 3, in which teachers created a divider at a T-cell count of 500).

The difficulty of describing portions of distributions using percents rather than counts, i.e., multiplicatively rather than additively, is one that has been documented by other research (Cobb, 1999), and we know that students and teachers struggle with this distinction using paper versions of dot plots. But two features in TinkerPlots™ highlight the issue. First, the presence of bins in comparing groups of unequal size provides an immediate need for multiplicative thought — how does one compare a bin in one distribution with the same bin in the other? How does one interpret the changing numbers as bins grow or shrink? Second, because TinkerPlots™ can provide the count and/or the percent of values in a bin, i.e., calculating the percent is no more difficult than counting the points, both values are equally available and the conversation about appropriate measures is quite likely to come up, as it did in our examples. We also note that looking at percents and NOT counts (as in Figure 7, in which there was a pie chart in each bin) can also lead one astray. After several months, most of the teachers reasoned multiplicatively most, but not all, of the time. Surprisingly, however, some teachers continued to slip back into the additive type of reasoning, although they could be "pulled back" out of it through conversation.

How do these results relate to Konold's taxonomy described above (Konold & Higgins, 2002; Konold et al., 2003)? The two relevant pieces of that taxonomy in this context are "classifier" and "aggregate." Those using a "classifier" view tended to view slices of the distributions out of context of the rest of the distribution, e.g. by comparing the number of students who make Bs between Amherst and Holyoke as a way to compare the entire two distributions. Even looking at these two bins as percents rather than counts doesn't solve the problem. Because the analysis ignores the distributions on either side of the bin of interest, it counts as a "classifier" view and does not answer questions about the entire distribution. On the other hand, making statements that take the entire distribution into account (e.g. 25% of this distribution is above 500 and 75% of the other distribution is above 500) is an example of "aggregate" thinking. So is using a measure of center such as the mean, but because of how easy it was for teachers to make other kinds of graphs, they rarely used measures of center.

Since TinkerPlots™ makes classifier and aggregate comparisons equally easy, the stage is set in a classroom for discussions of alternate analyses. Valuable perspectives can emerge from questions such as: Do two arguments about the same data set that rely on different representations always agree? What might be the relationship between an argument based on a classifier view and one based on an aggregate view of the same data set? Comparing data sets can also lead to new ideas: In what ways are the bins in the Amherst/Holyoke data similar to those some teachers created in analyzing the AIDS data? In what ways are they different? Because it is also possible to display measures of center and variability (e.g. box plots) on a graph, classroom discussions can include comparisons of using bins, cut points, and/or aggregate measures as a basis for analysis.

Teaching with a tool like TinkerPlots™ requires an in-depth understanding of the kinds of thinking the tool might engender and make visible. Once thinking is made visible, it can be discussed, challenged, and made more robust. But becoming aware of student thinking also raises new challenges for teachers, as these new ideas can be difficult to comprehend, their validity can be difficult to assess, and helping students work from them to more canonical ideas involves navigating complex and unexplored conceptual terrain. It is our experience that there is no substitute in teaching with a new tool for using it first as a learner, especially in a group in which alternate views of data and representation are likely to arise, as they certainly will in the classroom. Using a tool as a learner can

also help teachers experience the importance of complex data sets that pose particular challenges, e.g., unequal group sizes, and encourage the asking of more than one question.

We realize that many of our hypotheses and tentative conclusions are not "statistical" in nature, but we believe that the kind of study that follows a small group of teachers or students using new tools over two years can uncover new ways of thinking that a shorter and more controlled study could not. The teaching experiment model upon which this research is based (Ball, 2000; Cobb, 2000; Steffe and Thompson, 2000) uses a different set of methodologies and has different goals from a study that can be evaluated with a quantitative statistical analysis. A teaching experiment is based on a two-phase cycle that includes a development phase and a research phase. This cycle occurs multiple times during any teaching experiment, roughly after every session, in the course of planning for the next one. Teaching experiments make it possible to gather rich, multi-dimensional data that takes into account the interactions of individual students, the classroom social context, the resources available and the teacher's facilitation. A teaching experiment can identify multiple learning paths, as the methodology takes as a basic principle that individuals have different learning "stories."

The flip side of a teaching experiment's power to look deeply at a complex progression of events is that the information it provides is based on a small number of participants, e.g., a classroom of students, who are not guaranteed to be representative of a larger group. Based on our research, there are three categories of generalization that we feel are worthy of future study. 1) Generalizing to other groups. Do teachers not in the special subset of "those who were willing to collaborate with us" use TinkerPlots™ in different ways? How do middle school students use TinkerPlots™ to deal with variability? What difference would it make if we had just middle school teachers or just high school teachers in our seminar? 2) Generalizing to other interactive visualization tools. There are a few similar tools on the market now (TableTop™, Fathom™) and there are certain to be more. What aspects of these tools have similar affordances to TinkerPlots™? In what ways do they support different approaches to variability? 3) Generalizing to other professional development situations. We worked with this group of teachers for two years, and they were willing to stick it out for that long. What would happen in a shorter course? Do teachers use TinkerPlots™ differently if they have access to the software at home, not just in school? In addition to these issues of generalization, we have observed that some learners were more able to describe and discuss data as a sample. We have some preliminary evidence that this awareness may affect the representations that learners create and find compelling, but that is just the beginning of another complex story.

Our study explored the thinking of teachers and students as they grappled with, and tried to make intelligent use of, new software tools. An exploratory study of this kind, which involved researchers both as educators and observers of complex behaviors in a new arena, is bound to raise more questions than it answers, and we believe that is part of its value. We hope that our conclusions, while based on selective and at times incomplete evidence, can provide researchers and teachers with new ideas as well as with new research hypotheses regarding the role of new software tools in statistics education.

ACKNOWLEDGEMENTS

Support for this paper comes from the National Science Foundation (NSF) under grant REC-0106654, Visualizing Statistical Relationships (VISOR) at TERC, Cambridge, MA. The views expressed are those of the authors and do not necessarily reflect those of the NSF. A version of this paper was presented at The Third International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-3), Lincoln, Nebraska, USA, in July 2003. The authors are listed in alphabetical order; both contributed equally to the document.

Thanks to the teachers and students who have so graciously shared with us their thinking about these often difficult ideas. Thanks to our colleagues at TERC and SRTL and to the anonymous reviewers from SERJ whose questions and critiques have helped us improve upon earlier drafts of this paper. This work could not have been done at all without the contributions of Cliff Konold, designer of TinkerPlots™ and Bill Finzer, designer of Fathom™. Iddo Gal's comments as SERJ editor were

invaluable in helping us to continually sharpen our argument. And a special thanks goes to our incredibly adept research assistant, Camilla Campbell.

REFERENCES

Bakker, A. (2004). Design research in statistics education: On symbolizing and computer tools. Doctoral dissertation, Utrecht University.

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

Ball, D. L. (1991). Research on teaching mathematics: Making subject matter knowledge part of the equation. In J. Brophy (Ed.), *Advances in research on teaching* (Vol. II, pp. 1–48). Greenwich, CT: JAI Press.

Ball, D. L. (1993). With an eye on the mathematical horizon: Dilemmas of teaching elementary school mathematics. *Elementary School Journal, 93*(4), 373–397.

Ball, D. L. (2000). Working on the inside: Using one's own practice as a site for studying teaching and learning. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 365–402). Mahwah, NJ: Lawrence Erlbaum Associates.

Ball, D. L. (2001). Teaching with respect to mathematics and students. In T. Wood, B. S. Nelson & J. Warfield (Eds.), *Beyond classical pedagogy: Teaching elementary school mathematics* (pp. 11–22). Mahwah, NJ: Lawrence Erlbaum Associates.

Ball, D. L., Hill, H. C., Rowan, B., & Schilling, S. G. (2002). *Measuring teachers' content knowledge for teaching: Elementary mathematics release items, 2002*. Ann Arbor, MI: Study of Instructional Improvement.

Borko, H., & Putnam, R. T. (1995). Expanding a teacher's knowledge base: A cognitive psychological perspective on professional development. In T. R. Guskey & M. Huberman (Eds.), *Professional development in education: New paradigms & practices* (pp. 35–65). New York: Teachers College Press.

Case, R. (1978). Implications of developmental psychology for the design of effective instruction. In A. M. Lesgold, J. W. Pelligrino, S. D. Fokkema & R. Glaser (Eds.), *Cognitive psychology and instruction* (pp. 441–463). New York: Plenum Press.

Case, R. (1980). Intellectual development and instruction: A neo-Piagetian view. In A. E. Lawson (Ed.), *1980 AETS yearbook: The psychology of teaching for thinking and creativity,* (pp. 59–102). Columbus, OH: ERIC Clearinghouse for Science, Mathematics, and Environmental Education.

Clement, J. (2000). Analysis of clinical interviews: Foundations and model viability. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 547–589). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Cobb, P. (2000). Conducting classroom teaching experiments in collaboration with teachers. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 307–333). Mahwah, NJ: Lawrence Erlbaum Associates.

Cobb, P., Gravemeijer, K. P. E., Doorman, M., & Bowers, J. (1999). Computer Mini-tools for exploratory data analysis (Version Prototype). Nashville, TN: Vanderbilt University.

Cobb, P., McClain, K., & Gravemeijer, K. (2003). Learning about statistical covariation. *Cognition and Instruction, 21*(1), 1–78.

Duckworth, E. (1996). "*The having of wonderful ideas*" *and other essays on teaching and learning* (2nd ed.). New York: Teachers College Press.

Hancock, C. (1995). *Tabletop*. Cambridge, MA: TERC/ Brøderbund Software.

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

Harel, G., & Confrey, J. (Eds.). (1994). *The development of multiplicative reasoning in the learning of mathematics*. Albany, NY: SUNY Press.

Hill, H. C., & Ball, D. L. (2003). *Learning mathematics for teaching: Results from California's Mathematics Professional Development Institutes*. Ann Arbor, MI: University of Michigan.

Key Curriculum Press. (2000). Fathom™ Dynamic Statistics™ Software (Version 1.0). Emeryville, CA: Key Curriculum Press.

Konold, C. (2002). Alternatives to scatterplots. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics*, *Cape Town, South Africa*. [CDROM] Voorburg, The Netherlands: International Statistical Institute.

Konold, C., & Higgins, T. L. (2002). Highlights of related research. In S. J. Russell, D. Schifter & V. Bastable (Eds.), *Developing mathematical ideas (DMI): Working with data casebook,* (pp. 165–201). Parsippany, NY: Dale Seymour Publications.

Konold, C., Higgins, T. L., Russell, S. J., & Khalil, K. (2003). Data seen through different lenses. Unpublished manuscript, Amherst, MA.

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

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

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

Konold, C., Robinson, A., Khalil, K., Pollatsek, A., Well, A., Wing, R., & Mayr, S. (2002). Students' use of modal clumps to summarize data. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics*, *Cape Town, South Africa*. [CDROM] Voorburg, The Netherlands: International Statistical Institute.

Lamon, S. J. (1994). Ratio and proportion: Cognitive foundations in unitizing and norming. In G. Harel & J. Confrey (Eds.), *The development of multiplicative reasoning in the learning of mathematics,* (pp. 89–120). Albany, NY: SUNY Press.

Meletiou, M., & Lee, C. (2002). Student understanding of histograms: A stumbling stone to the development of intuitions about variation. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics*, *Cape Town, South Africa*. [CDROM] Voorburg, The Netherlands: International Statistical Institute.

Miles, M. B., & Huberman, M. (1984). *Qualitative data analysis.* Beverly Hills, CA: Sage Publications.

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

Mooney, E. S. (2002). A framework for characterizing middle school students' statistical thinking. *Mathematical Thinking and Learning, 4*(1), 23–63.

Mooney, E. S., Hofbauer, P. S., Langrall, C. W., & Johnson, Y. A. (2001). Refining a framework on middle school students' statistical thinking. Paper presented at the 23rd Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, Snowbird, UT.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage.

Rubin, A. (1991). Using computers in teaching statistical analysis: A double-edged sword. Unpublished manuscript, Cambridge, MA.

Rubin, A. (2002). Interactive visualizations of statistical relationships: What do we gain? In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics*, *Cape Town, South Africa*. [CDROM] Voorburg, The Netherlands: International Statistical Institute.

Rubin, A., & Bruce, B. (1991). Using computers to support students' understanding of statistical inference. *New England Mathematics Journal, 13*(2).

Rubin, A., Bruce, B., & Tenney, Y. (1990). Learning about sampling: Trouble at the core of statistics. Paper presented at the Annual Meeting of the American Educational Research Association.

Rubin, A., & Rosebery, A. S. (1988). Teachers' misunderstandings in statistical reasoning: Evidence from a field test of innovative materials. In A. Hawkins (Ed.), *Proceedings of the ISI Round Table Conference, Training Teachers to Teach Statistics*, Budapest, Hungary.

Russell, S. J., Schifter, D., Bastable, V., with Higgins, T. L., Lester, J. B., & Konold, C. (2002). *DMI Working with data: Casebook & facilitator's guide*. Parsippany, NJ: Dale Seymour Publications.

Russell, S. J., Schifter, D., Bastable, V., Yaffee, L., Lester, J. B., & Cohen, S. (1995). Learning mathematics while teaching. In B. S. Nelson (Ed.), *Inquiry and the development of teaching: Issues in the transformation of mathematics teaching* (pp. 9–16). Newton, MA: Center for the Development of Teaching, Education Development Center.

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

Schifter, D. (1997). *Learning mathematics for teaching: Lessons in/from the domain of fractions*. Newton, MA: Center for the Development of Teaching at Education Development Center, Inc.

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

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

Sowder, J. T., Philipp, R. A., Armstrong, B. E., & Schappelle, B. P. (1998). *Middle-grade teachers' mathematical knowledge and its relationship to instruction: A research monograph*. Albany, NY: SUNY Press.

Steffe, L. P., & Thompson, P. (2000). Teaching experiment methodology: Underlying principles and essential elements. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 267–306). Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

Watson, J. M., & Moritz, J. B. (2000). Developing concepts of sampling. *Journal for Research in Mathematics Education, 31*(1), 44–70.

JAMES K. HAMMERMAN & ANDEE RUBIN
TERC
2067 Massachusetts Ave.
Cambridge, MA, 02140
USA

REASONING ABOUT VARIABILITY IN COMPARING DISTRIBUTIONS

DANI BEN-ZVI
University of Haifa, Faculty of Education
dbenzvi@univ.haifa.ac.il

SUMMARY

Variability stands in the heart of statistics theory and practice. Concepts and judgments involved in comparing groups have been found to be a productive vehicle for motivating learners to reason statistically and are critical for building the intuitive foundation for inferential reasoning. The focus in this paper is on the emergence of beginners' reasoning about variation in a comparing distributions situation during their extended encounters with an Exploratory Data Analysis (EDA) curriculum in a technological environment. The current case study is offered as a contribution to understanding the process of constructing meanings and appreciation for variability within a distribution and between distributions and the mechanisms involved therein. It concentrates on the detailed qualitative analysis of the ways by which two seventh grade students started to develop views (and tools to support them) of variability in comparing groups using various statistical representations. Learning statistics is conceived as cognitive development and socialization processes into the culture and values of "doing statistics" (enculturation). In the light of the analysis, a description of what it may mean to begin reasoning about variability in comparing distributions of equal size is proposed, and implications are drawn.

Keywords: Variability; Comparing distributions; Statistical reasoning; Exploratory data analysis; Enculturation; Appropriation

1. SCIENTIFIC BACKGROUND

1.1. ENCULTURATION

Research on mathematical cognition in recent decades seems to converge on some important findings about learning, understanding, and becoming competent in mathematics. Stated in general terms, research indicates that becoming competent in a complex subject matter domain, such as mathematics or statistics, "may be as much a matter of acquiring the habits and dispositions of interpretation and sense making as of acquiring any particular set of skills, strategies, or knowledge" (Resnick, 1988, p. 58). This involves both cognitive growth and socialization processes into the culture and values of "doing mathematics or statistics". Many researchers have been working on the design of learning environments and teaching in order to "bring the practice of knowing mathematics in school closer to what it means to know mathematics within the discipline" (Lampert, 1990, p. 29). This study is intended as a contribution to the understanding of these processes in the area of Exploratory Data Analysis (EDA), focusing on reasoning about variability in comparing distributions.

One of the ideas used in this study is that of a process of *enculturation*, which is included in several recent learning theories in mathematics education (cf., Resnick, 1988; Schoenfeld, 1992). Briefly stated, this process refers to entering a community (or a practice) and picking up the community's points of view. The beginning student learns to participate in a certain cognitive and cultural practice, where the teacher has the important role of a mentor and mediator, or the *enculturator.* This is especially the case with regard to statistical thinking, with its own values and

belief systems and its habits of questioning, representing, concluding, and communicating. Thus, for *statistical enculturation* to occur, specific thinking tools are to be developed alongside collaborative and communicative processes taking place in the classroom.

1.2. RESEARCH ON VARIATION

Bringing the practice of knowing statistics at school closer to what it means to know statistics within the discipline requires a description of the latter. Based on in-depth interviews with practicing statisticians and statistics students, Wild and Pfannkuch (1999) provide a comprehensive description of the processes involved in statistical thinking, from problem formulation to conclusions. They suggest that statisticians operate, sometimes simultaneously, along four dimensions: investigative cycles, types of thinking, interrogative cycles, and dispositions. They position *variation* at the heart of their model of statistical thinking as one of the five types of fundamental statistical thinking.

Pfannkuch and Wild (2004) further explain the centrality of reasoning about variation in data inquiry problems:

> Adequate data collection and the making of sound judgments from data require an understanding of how variation arises and is transmitted through data, and the uncertainty caused by unexplained variation. It is a type of thinking that starts from noticing variation in a real situation, and then influences the strategies we adopt in the design and data management stages when, for example, we attempt to eliminate or reduce known sources of variability. It further continues in the analysis and conclusion stages through determining how we act in the presence of variation, which may be to either ignore, plan for, or control variation. Applied statistics is about making predictions, seeking explanations, finding causes, and learning in the context sphere. Therefore we will be looking for and characterizing patterns in the variation, and trying to understand these in terms of the context in an attempt to solve the problem. Consideration of the effects of variation influences all thinking through every stage of the [statistical] investigative cycle. (Pfannkuch & Wild, 2004, pp. 18–19)

According to Wild and Pfannkuch (1999), there are four aspects of variation to consider: noticing and acknowledging, measuring and modeling (for the purposes of prediction, explanation or control), explaining and dealing with, and developing investigative strategies in relation to variation. Reading and Shaughnessy (2004) suggest two additional aspects of variation that need to be considered; describing and representing. They claim that these six aspects of variation form an important foundation for statistical thinking.

Studies of reasoning about variation include investigations into the role of variation in graphical representation (Meletiou & Lee, 2002), comparison of data sets (Watson & Moritz, 1999; Watson, 2001; Makar & Confrey, 2004), probability sample spaces (Shaughnessy & Ciancetta, 2002), chance, data and graphs in sampling situations (Watson & Kelly, 2002), and variability in repeated samples (Reading & Shaughnessy, 2004). Hierarchies to describe various aspects of variation and its understanding have been developed by Watson, Kelly, Callingham, and Shaughnessy (2003) and by Reading and Shaughnessy (2004) in the context of repeated samples.

Noticing and understanding variability encompass a broad range of ideas. The basic form of variability in data is the variation of values *within* one distribution. Comparing distributions creates the impetus to consider other types of variability that exist *between* groups. Makar & Confrey (2004) discuss three different ways that teachers consider issues of variability when reasoning about comparing two distributions. They analyzed (1) how teachers interpreted variation *within* a group - the variability of data; (2) how teachers interpreted variation *between* groups - the variability of measures; and (3) how teachers *distinguished* between these two types of variation.

1.3. RESEARCH ON COMPARING DISTRIBUTIONS

Comparing groups provides the motivation and context for students to consider data as a distribution and take into account and integrate measures of variation and center (Konold & Higgins, 2003). At an advanced level, comparing distributions can stimulate learners to consider not only

measures of dispersion *within* each group, but comparisons of measures *between* groups, and hence to consider variation within the measures themselves (Makar & Confrey, 2004). Watson and Moritz (1999) suggest that comparing two groups provides the groundwork to the more sophisticated comparing of populations or two treatments in statistical inference. Without first building an intuitive foundation, inferential reasoning can become recipe-like, encouraging black-and-white deterministic rather than probabilistic reasoning.

There is some evidence however that the group comparison problem is one that students do not initially know how to approach and the challenge may remain even after extended periods of instruction. Students' difficulties may stem from the multifaceted knowledge necessary for comparing groups, such as understanding distributions (Bakker & Gravemeijer, 2004), representativeness (Mokros & Russel, 1995), and variability in data (e.g., Meletiou, 2002). Students also have difficulties in adopting statistical dispositions, such as tolerance towards variation in data, and integration of local and global views of data and data representations (Ben-Zvi & Arcavi, 2001; Ben-Zvi, 2002; Ben-Zvi, 2004).

Watson and Moritz (1999) observed two response levels in group comparison tasks completed by students during school years. In the first cycle, responses compared data sets of equal sizes, with or without success depending on the specific context. They did not recognize and/or did not resolve the issue of unequal sample size. In the second cycle, the issue of unequal sample size was resolved with some proportional strategy employed for handling different sizes.

There are a number of studies in which students who appeared to use averages to describe a single group or knew how to compute means did not use them to compare two groups (Bright & Friel, 1998; Gal, Rothschild & Wagner, 1990; Hancock, Kaput & Goldsmith, 1992; Konold, Pollatsek, Well, & Gagnon, 1997; Watson & Moritz, 1999). Konold et al. (1997) argue that students' reluctance to use averages to compare two groups suggests that they have not developed a sense of average as a measure of a group characteristic, which can be used to represent the group. Cobb (1999) proposes that the idea of middle clumps ("hills") can be appropriated by students for the purpose of comparing groups.

1.4. THE RESEARCH QUESTION

Based on these perspectives and studies, the following research question is used to structure the current study and the analysis of data collected: *How do junior high school students begin to reason about variability as part of an open-ended group-comparison task given in a rich and supportive classroom context?* Such a context involves a computerized environment, peer collaboration and classroom discussions, guidance of a teacher and curriculum-based tasks. The current study is different from some of the studies described above: It follows closely the dynamic behavior and discourse of two novice seventh grade students engaged with an EDA task. The students are observed within their classroom during an extended period of engagement with curriculum-based data investigation. A qualitative detailed analysis of the protocols is used, taking into account all kinds of actions, discussions and gestures within the situations in which they occurred. The goal is to trace the emergence of beginners' reasoning about variation in a comparing distributions situation, including the development of cognitive structures and the sociocultural processes of understanding and learning.

2. METHOD

Descriptions of the research setting, the statistics curriculum and the specific activity are followed by a profile of the students, technology used, and methods of data collection and analysis.

2.1. THE SETTING

This study took place in a progressive experimental school in Tel-Aviv, Israel. Skillful and experienced teachers, who were aware of the spirit and goals of the *Statistics Curriculum* (SC), taught

three classes. The SC was developed in Israel to introduce junior high school students (grade 7, age 13) to statistical reasoning and the "art and culture" of EDA (described in more detail in Ben-Zvi & Friedlander, 1997b; Ben-Zvi & Arcavi, 1998). The curriculum is characterized by the teaching and learning of mathematics using open-ended problem situations to be investigated by peer collaboration and classroom discussions using computerized environments (Hershkowitz et al., 2002). The design of the curriculum was based on the creation of small scenarios through which students can experience on their own, with limited teachers' guidance, some of the processes involved in the experts' practice of data-based enquiry. The SC was implemented in schools and teacher courses and subsequently revised in several curriculum development cycles.

The SC emphasizes student's active participation in organization, description, interpretation, representation, and analysis of data situations on topics close to the students' world, with a considerable use of visual displays as analytical tools (in the spirit of Garfield, 1995, and Shaughnessy, Garfield, & Greer, 1996). It incorporates technological tools for simple use of various data representations and transformations of them (as described in Biehler, 1993, 1997; Ben-Zvi, 2000). The scope of the curriculum is 30 periods spread over two to three months, and includes a student book (Ben-Zvi & Friedlander, 1997a) and a teacher guide (Ben-Zvi & Ozruso, 2001).

2.2. THE SURNAMES ACTIVITY

The *Surnames* activity, which is the focus of the current study, is the second full data investigation of the SC. It comes after an investigation involving the analysis of a time-series dataset with tabular data about Olympic 100 meters running times and a time plot of these data. The students are asked to compare the length of a set of surnames collected in their own class (35 Hebrew names) with a set of surnames from an American class that were given to them (35 English names). Equal sized data sets are used to simplify some aspects of the complex situation of comparing groups found in other studies (e.g., Gal, Rothschild & Wagner, 1990; Konold, Pollatsek, Well, & Gagnon, 1997), primarily students' difficulties with proportional reasoning. It was expected that the Surnames activity will support the development of beginners' reasoning about variability from the intuitive and simple to the more sophisticated and expert-like reasoning. The Surname data were given in a table (a part of it is presented in Figure 1).

Israeli Class (Hebrew names)				**American Class (English names)**			
Student's Number	First Name	Surname	Surname's Length	Student's Number	First Name	Surname	Surname's Length
1	מריה	אלקס	4	1	Kenneth	Auchincloss	11
2	מיכאל	אקרמן	5	2	Melinda	Beck	4
3	צפורה	בוסקילה	7	3	Edward	Behr	4
4	דולי	בטש	3	4	Patricia	Bradbury	8
5	רינה	בן שטרית	8	5	William	Burger	6
6	תמר	בריל	4	6	Mathilde	Camacho	7
7	איזבל	ברלין	5	7	Lincoln	Caplan	6

Figure 1. The upper part of the spreadsheet table displaying the raw data (There were 35 students in each class.)

In order to understand the analytic and interpretive challenge faced by the students, the two distributions are presented graphically in Figure 2. As a background, it is important to note that the variability between the two groups of names is in part due to differences in the structure of English and Hebrew. In modern Hebrew, as in Arabic and some other Semitic languages, words are often written without some vowels, making Hebrew words shorter than English words. Vowels are usually optional and if needed are written as diacritical marks under, within or above the letters, using dots and dashes which signify different types of vowels. These diacritical marks are not displayed in the second and third columns of Figure 1. There are additional cultural and historical factors that contribute to the variability in name length within and between the two language groups.

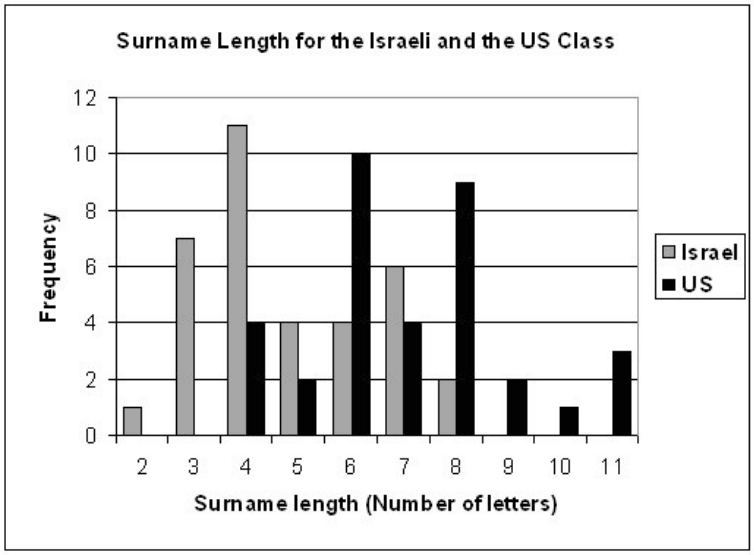

Figure 2. Double bar chart of the two surname groups

The whole Surnames activity took place during approximately three 90-minute lessons. Most of the time was spent on students' work in pairs in the computer lab, led by the student textbook. The teacher's interactions with the students were short and mostly occurred in reaction to their request for help. A session started with 5-10 minutes whole class introductory discussion and usually ended with a summary led by the teacher. In a preparatory lesson, students were asked: "What is the favorite shoe color and shoe size in your class? Compare the results to other seventh-grade classes". Students collected, organized, displayed and interpreted the data, compared the groups, and composed a summary report for a shoe company. Several statistical concepts and tools were informally introduced, or revisited, such as, statistical question and hypothesis, sample, categorical and quantitative variables, absolute and relative frequency, bar charts and frequency table.

In the following lessons, which are the focus of this report, three methods were offered by the curriculum-based materials to compare distributions: (a) absolute and relative frequency distributions presented in tables; (b) basic measures of variation and center, such as range, mode, mean, and median; and (c) graphical representations, such as a double bar chart. These statistical methods and tools were introduced to help students in describing and interpreting the surnames data and the variability in it, searching for trends and drawing conclusions on comparing the two groups. The purpose of the activity was to set the stage for students to consider data as a distribution and provide many opportunities to notice, acknowledge, intuitively deal with, and describe the variability within and between distributions.

2.3. PARTICIPANTS

This study focuses on two students, *A* and *D*, who were above-average ability students (grade 7, age 13), very verbal, experienced in working collaboratively in computer-assisted environments, and willing to share their thoughts, attitudes, doubts, and difficulties. They agreed to participate in this study, which took place mostly within their regular classroom periods and included being videotaped and interviewed (after class) as well as furnishing their notebooks for analysis. While not necessarily representing their classmates, verbal and able students provide a better opportunity for collecting valuable and detailed data on their actions, thoughts and considerations.

When they started to learn this curriculum, *A* and *D* had limited in-school statistical experience. However, they had some informal ideas and positive dispositions toward statistics, mostly through exposure to statistics jargon in the media. In primary school, they had learned only about the mean and the uses of some basic diagrams, such as bar and pie charts. Prior to, and in parallel with, the learning of the SC they studied beginning algebra based on the use of spreadsheets to generalize numerical linear patterns (Resnick & Tabach, 1999).

The students appeared to engage seriously with the curriculum, trying to understand and reach agreement on each task. They were quite independent in their work, and called the teacher only when technical or conceptual issues impeded their progress. The fact that they were videotaped did not intimidate them. On the contrary, they were pleased to speak out loud, address the camera explaining their actions, intentions, and misunderstandings and share what they believed were their successes.

2.4. TECHNOLOGY

During the experimental implementation of the curriculum a spreadsheet package (Excel) was used. Although Excel is not the ideal tool for data analysis (Ben-Zvi, 2000), there are several reasons for choosing this software. Spreadsheets provide direct access that allows students to view and explore data in different forms, investigate different models that may fit the data by, for example, manipulating a line to fit a scatter plot. Spreadsheets are flexible and dynamic, allowing students to experiment with and alter displays of data. For instance, they may change, delete or add data entries in a table and consider the graphical effect of the change or manipulate data points directly on the graph and observe the effects on a line of fit. Spreadsheets are adaptable by providing control over the content and style of the output. Finally, spreadsheets are common, familiar, and recognized as a fundamental part of computer literacy (Hunt, 1995). They are used in many areas of everyday life, as well as in other domains of the mathematics curricula, and are available in many school computer labs. Hence, learning statistics with a spreadsheet helps to reinforce the idea that this is something connected to the real world.

2.5. DATA COLLECTION AND ANALYSIS

A diverse body of data was collected to study the effects of the new curriculum. The behavior and reasoning of the two students on which the present study focused was analyzed using lengthy video recordings of whole class sessions, classroom observations, interviews, and students' notebooks and research projects. In addition, observational data and summative assessment data were also collected for the whole class to support other research objectives, but are beyond the scope of this paper.

The analysis of the videotapes was based on interpretive microanalysis (see, for example, Meira, 1991): a qualitative detailed analysis of the protocols, taking into account verbal, gestural and symbolic actions within the situations in which they occurred. The goal of such an analysis is to infer and trace the development of cognitive structures and the sociocultural processes of understanding and learning.

Two stages were used to validate the analysis, one within the SC researchers' team and one with four researchers in science education, who had no involvement with the data or the SC (triangulation in the sense of Schoenfeld, 1994). In both stages the researchers discussed, presented, and advanced and/or rejected hypotheses, interpretations, and inferences about the students' cognitive structures. Advancing or rejecting an interpretation required: (a) providing as many pieces of evidence as possible (including past and/or future episodes, and all sources of data as described earlier) and (b) attempting to produce equally strong alternative interpretations based on the available evidence. In most cases the two analyses were in full agreement, and points of doubt or rejection were refuted or resolved by iterative analysis of the data. In the presentation of transcripts, comments in block parentheses are clarifications suggested by the author, and were verified by a triangulation process.

3. RESULTS: STUDENTS' DEVELOPMENT OF REASONING ABOUT VARIABILITY

This paper describes how A's and *D*'s novice views slowly changed and evolved towards an expert perspective while comparing two data sets of the same size. The focus is on how they began to notice and acknowledge variability in the data and make use of special local information in different ways as stepping-stones towards the development of global points of view of describing and explaining the variability between the groups. The study identifies seven developmental stages of their reasoning about variability (Figure 3).

Stage 1	On what to focus: Beginning from irrelevant and local information.
Stage 2	How to describe variability informally in raw data.
Stage 3	How to formulate a statistical hypothesis that accounts for variability.
Stage 4	How to account for variability when comparing groups using frequency tables.
Stage 5	How to use center and spread measures to compare groups.
Stage 6	How to model variability informally through handling outlying values.
Stage 7	How to notice and distinguish the variability within and between the distributions in a graph.

Figure 3. The seven suggested stages through which the two students progress

Stage 1. On What to Focus: Beginning from Irrelevant and Local Information

When the teacher introduced the whole class to the *Surnames* problem situation, she asked the students to hypothesize about interesting phenomena regarding names in general, without first providing them with any data. After a brief discussion about students' intuitive hypotheses, the teacher focused the discussion on name length in various cultures and countries, and presented the main task: Compare the surname length of the Israeli and the American groups. The teacher considered some sample quick responses (e.g., "*American surnames are longer than Israeli surnames*", "*They are about the same*") as an indication that the students had enough familiarity with the context of the task in order to engage meaningfully with the data. When the introduction was over, *A* and *D* moved to the school computer lab to work on the Surnames activity. Their work was guided by a list of questions that appeared in the Student Workbook which was part of the SC.

After *A* and *D* added the names of their classmates to the Excel table (a part of it presented in Figure 1), they started working on the first question in their Workbook, "*Look at the table and suggest a research question about length of surnames*." The raw data, i.e., names, were displayed in a table on the computer screen. After a short discussion they agreed on posing the question, "*Which of the two countries has longer names?*" This initial focus on finding the "winning" group resembles the type of questions suggested in the introductory whole class discussion and was typical of students' questions in the experimental classes. This formulation, deterministic in nature and ignoring the complexity involved in comparing groups, is not surprising at this beginning stage of working on a complex data-analysis task.

In the second question, students were asked to formulate a hypothesis regarding interesting phenomena in the data. The question, which was proposed to 'push' students to look at the data and consider patterns and variability, provoked the following exchange between *A* and *D*. (The row numbers in the transcripts are provided to assist later in referring to specific sentences.)

1	*A*	We have to phrase now a hypothesis regarding interesting phenomena in the data.
2	*D*	Interesting phenomena, interesting phenomena. O.K., we should find interesting phenomena. We'll find interesting phenomena. [Reads the question again] "Formulate a hypothesis about interesting phenomena in the length of surnames". I didn't understand what it exactly means.
3	*A*	O.K., lets skip this [question], since we don't have anything interesting at hand. We may shortly find something.
4	*D*	I don't think we should skip this, we'll simply ask what the precise intention is. I didn't really understand: Shall we hypothesize about 'Mc's'? [There are three surnames in the American class, beginning with the letters Mc, such as McDaniel.] No! I don't understand. [Laughing] This isn't funny. I'll ask Michal [their teacher] to come and help us.

Their remarks indicate that questions like "phrase a hypothesis regarding interesting phenomena in the data" may encounter an initial inability to focus attention on relevant (even informal) aspects of the data. *A* and *D* seemed to be unable to make full sense of the intention of the question and its formulation. Their focus on irrelevant features of the data, or their inability to focus on anything at all [row *3*], is similar to their reaction at the beginning of the first problem situation in the SC–*Olympic*

Records (analyzed in detail in Ben-Zvi & Arcavi, 2001). In both activities, they were aware that their observations, such as names beginning with Mc, might not be relevant. They somehow recognized what not to focus on, but were uncertain about what may qualify as 'interesting phenomena' in this context, or how to reply to such questions, and finally requested the teacher's assistance to help them overcome this difficulty.

In the above brief discourse the students did not notice global features of the data and the variability within it. Their initial local focus on what they saw as outstanding regularity in the data (the three "Mc" surnames) seems to restrict them from observing the distributions as a whole. Interestingly, this phenomenon was already observed when these students worked on their first activity of the SC (see section 2.2). There, *they* were similarly attentive to the prominence of "local deviations" in data and this appeared to keep them from creating more global interpretations of data. Only after the following teacher intervention were they able to start focusing on relevant information, taking into account the variability in the data.

Stage 2. How to informally describe the variability in raw data

When *A* and *D* requested the teacher's help in answering the hypothesis task, the following dialog took place.

5	*A*	[Asking the teacher] What does it mean?
6	*D*	What does it mean to "phrase a hypothesis about interesting phenomena"?
7	*A*	That there are many names beginning with 'Mc'?
8	*T*	About the length of surnames. OK?
9	*A*	What is 'interesting phenomena'?
10	*T*	Are there no interesting phenomena in the data?
11	*A*	[Cynically] It's very interesting that there is a Michael…
12	*T*	You are asked about length!
13	*D*	About length … An interesting phenomenon is that there is a [counting letters in the Hebrew name Levkowitz] 1, 2, 3, 4, 5, … [7] letter name here and a 4 there [Cose in the American class].
14	*T*	OK. You suggest that there are very short names and very long ones.
15	*A*	Do we have to compare?
16	*D*	So what's the hypothesis?
17	*T*	I don't know [what the hypothesis is]. First, it's a phenomenon. What do you think? Are there many long or many short [surnames]?
18	*A*	There will be a lot more of the long in USA.
19	*D*	More long than short.
20	*T*	OK. You have a hypothesis: In the USA…
21	*A*	But what is long, and what is short?
22	*T*	That's a different question.
23	*A*	What should we write?
24	*D*	Perhaps longer than this? Or…
25	*A*	What name is considered long?
26	*T*	OK. Longer than this – that's a comparison. When you compare these groups, you say – I expect that there will be so and so here… That's comparing two groups. That's all right.

The students were uncertain about the intention of the question ("phrase a hypothesis") as well as the meaning of the phrase "interesting phenomena". The fact that a particular research question (comparing the two groups in terms of surname length) had been introduced at the beginning of the activity did not help them to focus and they seemed to be overwhelmed by the complexity of the data. Their initial observations are irrelevant and local (Mc's, Michael). It seems that there are three factors interacting to produce the students' inability to proceed: (a) the lack of understanding of the intent of the question, (b) the lack of understanding of the phrase "interesting phenomenon", and (c) the complexity of the data. These factors played a role in causing confusion in other parts of the transcripts of these students (cf. Ben-Zvi & Arcavi, 2001).

The teacher's initial help consisted of calling their attention twice to the investigated variable, namely, the length of a surname. Only her second trial [row *12* in the transcript above] pushed *D* to compare the surname length of two students (one from each class) [*13*]. Thus, he began focusing on the correct variable and noticing one aspect of the variability in the data, but in a very local way. The teacher accepted his answer as being in the right direction, and suggested a generalization of his local observation [*14*]. This intervention represents a generalization 'jump' by the teacher not reflected in the students' previous comments. She then nudged them to quantify the variability in the data in a simple way [*17*].

In response to the teacher's direct question, the students suggested that the long surnames are more frequent in the USA group [*14–15*]. It is hard to determine at this point if *A* considered only the variability within the American group, or the variability between the groups. Whichever interpretation is taken here, this initial consideration of variability later became the foundation on which *A* and *D* developed an informal model of the variability within, as well as between, the two groups. The students' first attempts to describe the variability in the data by comparing long and short names raised a new concern about the borderline between long and short names [*21*], which was not resolved at this stage, and may be the beginning of an attempt to handle variability by grouping the data. The interaction with the teacher closed with her recommendation to focus on comparing groups.

Stage 3. How to formulate a statistical hypothesis that accounts for variability

The above interaction with the teacher helped the students to re-focus and propose a hypothesis. The following dialogue between *A* and *D* took place immediately after the teacher left them.

27	*D*	Our hypothesis about interesting phenomena in the length of surnames is: In the USA, surnames will be...
28	*A*	Will be longer...
29	*D*	Longer than in Israel…
30	*A*	Usually than in Israel...
31	*D*	Usually, not always, usually.
32	*D*	Let's see, we have Levkowitch here [in the Israeli class] and Cose there [in the American class] – that's different.
33	*A*	Enough, enough, come on.
34	*D*	OK, never mind.
35	*A*	So, in the USA... the surnames…
36	*D*	Will be usually longer.
37	*A*	Very nice!

After the previous discussion with the teacher, the students were able to formulate a sensible hypothesis regarding the comparison between the two groups that took into account the variability in the data. They began with a deterministic proposal for a rule, 'surnames in the USA are longer than in Israel'. However, they noticed immediately that this assertion does not take full account of the situation presented by the data, and decided that variability should be included in their description by adding the constraint "*usually, not always*" to the rule. Understanding that some surnames can "behave differently", i.e., deviate from a general rule they formulated, can be considered an important step in the development of their acceptance of the existence of and tolerance to variability. In other words, they began to adopt the statistical perspective of trends that are generally true, but still have exceptions.

This new understanding is evident in *D*'s provision of an "opposite example", an Israeli name that is longer than a USA name, to show that the 'rule' holds even if there are opposite cases. *D* suggested this same example in the previous discussion with the teacher. While at that time it limited his ability to formulate a general hypothesis and view the data globally, here it is an expression of comfort with global views of the data that include variability. Hence, this opposite example, which derailed *D* from being on the right track on the first occasion, helped him adopt a statistical view of variability at this subsequent time.

Why might the students have initially focused on deterministic relationships between the variables and paid special attention to the unusual case? A possible explanation for their perspective can be found in their short-term learning history. *A* and *D* used spreadsheets in their algebra studies (immediately before they started to learn the EDA unit), to explore patterns, generalize, model mathematical problems, create and use formulae, and draw tables and graphs. Most of the tables investigated were linear correspondences between two sets of values. The students were accustomed to generating tables with the spreadsheet by 'extending' the pattern of constant differences between adjacent cells through the act of 'dragging' a pair of cells to duplicate this difference to the rest of the cells in the column resulting in long tables with clearly defined patterns. Using the same exploratory learning environment may have evoked for them the same deterministic nature of the relationship between variables found in algebra, which they incorrectly applied in statistics in order to make sense of data. Thus, their first focused observations referred to what was salient to them and a familiar part of their practices, the 'differences' between adjacent data entries not being constant. The only regularity they found in the data was a set of three Mc names. Maybe they implicitly began to sense that the nature of these data in this new area of EDA, as opposed to algebra, is disorganized, and it is not possible to capture it in a single deterministic formula, e.g., the previous "*Usually, not always*" comment.

At the end of this episode the two students seemed very satisfied with their answer. However, it was hard to appreciate at this stage how fragile their current understanding was. Additional difficulties with their abilities to acknowledge, explain, describe and deal with the variability in data in the context of this "noisy" and complex data situation unfolded in later stages of their work.

Stage 4. How to account for variability when comparing groups using frequency tables

After the students formulated a research question and hypothesis they were introduced by the student textbook to different concepts related to frequency in the context of the surnames investigation: frequency, relative frequency, and creating univariate frequency tables using spreadsheets. At this stage, *A* and *D* worked smoothly with the software and tasks, explaining every step and overcoming technical and conceptual hurdles. The following dialogue took place when they completed the production of two univariate frequency tables and were asked to use them to compare the two groups. See Figures 4 and 5, which are recreations of actual displays students generated on their own.

Israeli class		
Surname's length	Frequency	Relative frequency (%)
2	1	3
3	7	20
4	11	31
5	4	11
6	4	11
7	6	17
8	2	6
Total:	35	100%

Figure 4. Frequency table of surname's lengths in the Israeli class

American class		
Surname's length	Frequency	Relative frequency (%)
4	4	11
5	2	6
6	10	29
7	4	11
8	9	26
9	2	6
10	1	3
11	3	9
Total:	35	100%

Figure 5. Frequency table of surname's lengths in the American class

38 *D* [Reads the task] Use the frequency tables that you generated to compare the surnames' length in the two countries… Emm… They [the American surnames] are really a little bit longer. In the USA there are no 2 or 3-letter names…

39 *A* Yes. And in Israel…

40 *D* … since they [the 2 or 3-letter names] are a bit short.

41 *A* The table [Figure 4] starts from…

42 *D* From 2 [letters] to 8 [letters].

43	*A*	The [Israeli] surname length is from 2 to 8… And in the USA they're from 4 to 11… In other words, in the USA 2 or 3-letter names are not considered at all.
44	*D*	They're considered, but there are simply none.
45	*A*	There are none, or there is exactly one in the whole USA, something like that… And in Israel, names with 9, 10, and 11 letters are not considered, because there are none.
46	*D*	Because they [American names] have vowels. For example, Raz, Itzik Raz [a student in their class]: Here [in Hebrew] it's R and Z, and there [in English] it's R, A, and Z – three letters, did you understand?
47	*A*	In Israel, names with 9, 10, and 11 letters are not at all considered, because there are none. There may be one or two all over the country, yes, yes.
48	*D*	Like Levkowitch.
49	*A*	So, for example, we see that names with 8 letters are 6% in Israel.
50	*D*	There – they are 26%.
51	*A*	In the USA they are 26%.
52	*D*	20% more.
53	*A*	20% more, and it's a lot more, and…
54	*D*	A lot more, interesting, lovely… Actually, emm… just a second… That's exactly all I'm saying…
55	*A*	I assert that in the USA there are more… the names…
56	*D*	There are longer names, right.
57	*A*	Longer according to the comparison between these tables [Figures 4 & 5]. It may not be certain, but at least according to these tables… So, in the USA table, there are no 2 and 3 letter-names while there are 9, 10, and 11, but none in Israel. This means that the names are longer. [Writing this conclusion in his notebook.] Now, we also see here that in Israel, there are many more 4-letter names, which is considered pretty short.
58	*D*	Having a 4-letter name is the coolest matter in Israel.
59	*A*	So maybe because of that, there are more of those [surnames] in Israel, and in the USA – the names are longer. Therefore there aren't many names with 4 letters there. I brought up the 4 letters just as an example.

The students were faced with an unfamiliar and complex situation, presented in two separate frequency tables that included many values (Figures 4 & 5). Their purpose was to find ways to justify their hypothesis that surnames in the USA are usually longer than in Israel using the two frequency tables they had just created. On their own, they constructed a comprehensive argument, consisting of the comparison of two kinds of "special" values within the distributions: disjoint edge values – present in one distribution and absent from the other (and vise versa), and common edge values – the first and the last common values of the two distributions.

They began their argument by looking at the distributions' edges, moving from the lowest to the highest edge, and the range of values in between. *D* used the left "tail" (the shortest surnames in Israel that are missing in the USA group) as a justification for the claim that American surnames are "*a little bit longer*" [*38*]. They continued by noticing the different ranges of the groups; however, they did not make explicit use of them as measures of dispersion [*41–43*]. Then *A* argued symmetrically about the right "tail" of the USA distribution that is missing in Israel. While this opposite symmetry between the distribution edges seems to strengthen their confidence in the claim that the USA surnames are longer, it does not help them see the horizontal shift between the two generally-similar distributions.

Once the disjoint values were considered, the students moved on to compare the frequencies of the neighboring values, namely the last and the first common values of the distributions (8 and 4-letter names respectively). *A* suggested that the large differences in the relative frequencies of these values provided additional support to their hypothesis. They also informally acknowledged that 4-letter surname is the 'mode' in Israel [*58*]. These comments may represent first steps towards understanding density in a distribution.

A and *D* integrated contextual knowledge to support their understanding of, and in order to account for, the variability in the data. First, *D* suggested a causal explanation to account for the group differences, namely the use of vowels in English versus diacritical symbols in Hebrew. He also provided an example of one Israeli surname Raz, which has three letters in English but only two in Hebrew [*46*]. *A* further speculated that their sample implied the rarity of very short and very long

surnames in the USA and the Israeli populations respectively [*47*]. *D* supported him bringing up his frequently mentioned example of Levkowitz, a relatively long Israeli surname in their class. In these actions, *A* and *D* were trying to synthesize statistical and contextual knowledge to draw out what can be learned from the data about the context of the problem. The context of the problem supports their statistical reasoning by providing reasonable explanations to the emergent patterns in the variation. At the end of this dialogue they wrote the following synthesis in their notebooks.

A "~~In the USA, the names are longer than in Israel~~. [This sentence was written and later erased by A.] In the American table, there are no names with 2 and 3 letters, and there are names with 9, 10, 11 (none in Israel). In Israel, short names are more frequent; In the USA, the long names are more frequent."

D "In the USA, the names are longer than in Israel (according to the tables). In the American table, there are no names with 2, 3 letters, and there are of 9 to 11."

Arriving at a general conclusion was not a straightforward process for both students; however, they seem to be in different positions. *D,* without much doubt, accepted that the conclusion "*In the USA, the names are longer than in Israel*" captured the essence of the situation, and was less disturbed by the presence of outlying values, or irregular patterns in the data. In contrast, *A* struggled more with the variability presented in the data, and was more attentive to the prominence of "local deviations", which kept him from dealing more freely with global views of data. This could have been the reason for his erasing the general conclusion in his written summary. On the other hand, the rest of his conclusion is a beginning step to modeling variability and conceptualizing the use of 'density' in comparing distributions.

Stage 5. How to use center and spread measures to compare groups

In the second part of the *Surnames* activity the students were introduced to basic statistical measures of center (mode, mean, and median), spread (range) and outliers. They used the computer to find the statistical measures of the two groups and organized them in a table. See Figure 6 which is a recreation of the actual display the students generated. The next question was to use these measures to compare the groups. The students were uncertain how to answer the question and asked for help. After the teacher approved one answer as being in the right direction, *A* and *D* started to interpret the table.

Statistical Measures	Israeli Class	American Class
Number of Students	35	35
Mode	4	6
The maximal value	8	11
The minimal value	2	4
Range	6	7
Mean	4.83	7.06
Median	4	6
Outlying values	2, 8	5, 9, 10

Figure 6. Statistical measures of the two classes (The correct median of the USA group is 7. For the outliers, the students chose values with minimal frequency.)

Using the statistical measures table that they generated (Figure 6), the students started comparing the groups by noticing that both the maximal and the minimal values of the Israeli group are smaller than those of the American group. However, they erroneously concluded that the range is also smaller since the two extreme values are smaller in the Israeli names. While the range does happen to be smaller, it is not for the reason stated. This shows a misinterpretation on the part of the students. Once they noticed that the mean and the median also behaved in a similar way, they inferred that all the statistical measures of the Israeli distribution are smaller than those of the USA distribution. In spite of their fluent work at this stage, their actions seem to be merely procedural, missing both the meaning of measures as representative numbers (Mokros & Russel, 1995), and the distinction between center and spread measures.

Stage 6. How to model variability informally through handling outlying values

Dealing with information in the last row of the measures table (Figure 6) initiated the following dialogue about outliers.

60	*D*	But in the outlying values…
61	*A*	In fact here it's [different than the rest of the measures]… You expect that in Israel the outlying values will be higher [larger] than in the USA, since there are less high [long surnames in Israel]. But in fact you see here that in Israel the outlying values are not so high [large].
62	*D*	I am confused now, I don't understand. Not correct, because if your data…
63	*A*	If everything in Israel is smaller, then you would expect that the outlying values, yes, will be high [large] numbers, since there are few of them; and in the USA, the outlying numbers – will be lower [smaller], since there are few of the low [short surnames].
64	*D*	Yes, but this is not correct.
65	*A*	But in fact in the USA also - the high [large surnames] are the outliers.
66	*D*	9 and 10.
67	*A*	Right, 10 and 9 are outliers, but 11 is really high [long].
68	*D*	Correct.
69	*A*	Well, let's not write about that.

So far, the comparing of the two groups using statistical measures had been a straightforward and monotonous task. However, the outliers in the last row of the measures table presented a new challenge to the students: how to compare sets of numbers (2 and 8 in Israel vs. 5, 9, and 10 in USA) that had no trivial pattern and meaning. Furthermore, *A*'s pre-conceptualization of outliers as unusual and least frequent values in a distribution made him predict that the outliers in Israel would be only the long surnames since the Israeli surnames tended to be short (and vice versa in the USA distribution).

A seems to deal with distributions' variability with a plain dichotomous model. In his mental model, he divides the distributions to two groups: The short surnames that include the majority of the Israeli values, and the long surnames - the minority (and vice versa in the USA). This model appears to have helped him deal, describe and quantify the variability by reducing the 'noise' within the distributions. He consequently predicted that the variability between the groups would be also straightforward [*61*]. Once the students realized that the outliers were telling them a conflicting, more complex 'story' of the variability in the data, they did not find an alternative explanation and gave up on the resolution of the conflict.

It appears that having to deal with the outlier as a concept (i.e., a principled class of observations, not just some specific data points) contributed to the complexity of the students' conceptual task and understanding at this stage. A few minutes before the above dialogue took place, they came across outliers and chose to define them as "*the highest and the lowest values*". The meaning of the Hebrew word for outlier is "exceptional or unusual" and may have influenced their definition choice. Thus, from their perspective, the modal value was also an outlier. The teacher's explanation that outliers are individual data points that fall outside the overall pattern of the distribution made them abandon the mode as an outlying value, but left them with the view of outliers as merely the least frequent values.

Through their dealing with the outliers, the students presented a simplistic view of the distributions in order to handle the variability in the data. In their model, resembling a skewed distribution, the majority of the distribution concentrates in one interval, while the less frequent values, the outliers, are positioned in a disjoint interval. This model helped them to present clearly the difference between the distributions, which followed opposing patterns. In their view, the selection of outliers is based on low frequencies, meaning they are exceptional, since they are rare. In that respect, the students' consistent use of "high" and "low" to describe the "long" and "short" surnames in all the dialogues can be attributed to their focus on the variability in frequencies and not only to a careless use of language.

Stage 7. How to notice and distinguish the variability within and between the distributions in a graph

In the third and final part of the activity, the students were guided to generate graphical displays of the data and were asked to use them to compare the distributions. The following dialogue took place after they created a double bar chart of the two groups (similar to the graph displayed in Figure 2).

70	*D*	[Reading the task] Use the graph you generated (Figure 2) to describe the emerging trend in the surnames' length of the two countries.
71	*A*	Let's see: The USA… usually… no… hold on…
72	*D*	It seems that it's a lower trend in the USA.
73	*A*	Not low, it seems about the same in the graph.
74	*D*	Aha… No, higher trend.
75	*A*	Hold on, the USA…
76	*D*	Since you do not compare this to that, but rather this to that.
77	*A*	[Cynically] Really!
78	*D*	All right. [Unclear] … seven.
79	*A*	So it's higher here, it's higher here, here, here, and it's higher here; but in Israel it's higher here, here, here, and here.
80	*D*	And here.
81	*A*	And here.
82	*D*	They balance each other.
83	*A*	Look, the advantages [height differences] are bigger in Israel. No, not always. Let's ask someone [a teacher] what it means.
84	*D*	I know what it means.
85	*A*	What?
86	*D*	It means that the emerging trend is…
87	*A*	But it is not equal. Look, we said that the USA is longer… The USA leads in 8, 9, 10, and 11, while Israel leads only in 2, 3, 4, 5, and…
88	*D*	We said that the USA names are longer, what's the big deal?
89	*A*	That's right. So, the USA leads in the longer names. That's also not a big deal since 2 was not considered at all in the USA, while 11 was not considered at all in Israel.
90	*D*	What's the big deal? They were not considered because there are none.
91	*A*	OK, but…
92	*D*	They did not ignore data… It appears that in Israel the lengths of the lower names are…
93	*A*	No…
94	*D*	The length of the names
95	*A*	In Israel… In Israel…
96	*D*	The lengths of the lower names are…
97	*A*	No. In Israel, the lengths of names with fewer letters have a higher frequency, but in the USA, the lengths with… [having difficulties to complete the sentence]
98	*D*	I know how to formulate this. Write down.
99	*A*	No. I first want to hear what you have to say.
100	*D*	OK. In Israel, the frequency of the names with low number of letters…
101	*A*	Relatively low.
102	*D*	… is higher than in the USA.
103	*A*	Just a second, low – let's say smaller than 5.
104	*D*	Let's assume so. …is higher than…
105	*A*	No. But there is also one exception here.
106	*D*	The frequency is higher than in the USA.
107	*A*	But there is also one exception here.
108	*D*	[Angrily shouting] OK, it's in general! It's a general trend! It's not the trend for the exceptional one.
109	*A*	[Surprised by D's reaction] Buu …
110	*D*	OK. On the other hand, in the USA, the trend… the frequency of the long surnames is relatively higher than in Israel.

Although the students are familiar with generating and interpreting bar-graphs, handling this particular double bar-graph (Figure 2) is a complex task for them. Their challenge is figuring out the graph and understanding the variation embedded in the data. At first, the students provided conflicting interpretations of the graph; their rather unclear statements [72-73] are initial attempts to find one global description that accounts for variability by summarizing the difference between the bars in the two groups. This attempt can be considered a progress in comparison to their previous interpretations of graphs in the SC, which were mostly local, focusing on one or more individual values within the distributions (Ben-Zvi & Arcavi, 2001). *D* suggested that their disagreement arose from their different ways of reading the graph: 'horizontal' reading – comparing values, vs. 'vertical' reading – comparing heights of bars (density, frequency).

The students then began focusing on comparing the heights of adjacent bars from the two groups. Based on a method *A* suggested for summarizing the differences between the groups [76-79], they counted how many times the bar of one group was higher than the bar of another group for each surname value on the X-axis. For example, if for a surname length of 6 letters the bar for the Israeli group was of height 4 and for the US group of height 10, then the US group was "winning" there. However, this led them to an impasse: the number of "winning" Israeli and American bars was equal [*82*]. A second trial to compare the height differences between adjacent bars also proved fruitless.

Only when they began focusing on the *location* of the "winning" bars of each group, did they realize that the American bars are higher than the Israeli bars for the long names, while the Israeli bars are higher for the short names. Thus, they reduced the problem of comparing each pair of bars to comparing two subgroups, the relatively short and long surnames. Their previous success, in the frequency table task, in handling the variability between the groups by dividing the distributions to two groups seems to have helped the students out of impasse also here. This informal comparing method resembles Cobb's (1999) finding that the idea of middle clumps ("hills") can be appropriated by students for the purpose of comparing groups.

However, *A* was not completely satisfied with the above realization and was particularly concerned [*103*] about the distinction between short and long names. This issue, which worried him also at the beginning of the activity [*21*], was triggered here by the lack of clear-cut borderline between the groups: 5 and 7-letter names are more frequent in Israel and the 6-letter names are more frequent in the USA (see Figure 2). While *A* could not ignore the presence of this deviation in favor of a global summary of the variability between the groups, *D* was not disturbed by the 'noise' in the data. He claimed that their comparison is general and therefore they must ignore the one exception [*108*].

They requested the teacher's approval before they wrote a summary in their notebooks: "*The emerging trend is that the frequency of relatively short names (up to 5 letters) is higher in Israel than in the USA, but the frequency of relatively long names is higher in the USA than in Israel.*" Thus their final description of the variability between the groups was based on comparing the frequencies of two subgroups ignoring the deviation from the trend in the center.

4. DISCUSSION

This study was undertaken to contribute to our understanding of the process through which students develop ways to reason about variability within and between distributions. The study examined the first steps of two students who worked on a group-comparison task in a rich technology-based environment. In this environment, as happens in regular classes, students' work and intuitions are supported by formal curricular materials and ongoing instructional activities. The results illustrated several aspects, discussed below, of students' emerging understanding of variability in comparing groups and the role of supporting factors in that process, in particular the teacher's role. Conclusions and implications are discussed further below.

4.1. STAGES IN DEVELOPMENT OF REASONING ABOUT VARIABILITY

A and *D* started by trying to make sense of general questions normally asked in EDA tasks. Their learning trajectory included coming up with irrelevant answers and feeling an implicit sense of

discomfort with them, asking for help, getting feedback, trying other answers, working on a task even with partial understanding of the overall goal, and confronting the same issues with different sets of data and in different investigation contexts. This problem-solving process is consistent with several other research findings (see, for example, Moschkovich, Schoenfeld, & Arcavi, 1993; Magidson, 1992): novices may be either at a loss (when asked these kinds of questions) or their perceptions of what is relevant are very different from the experts' view.

When looking at raw data (stages 1-2), the students initially did not notice global features of the data and the variability within them. Their initial focus on what they saw as outstanding regularity in the data, the three "Mc" surnames, was based on attention to local features and seems to have restricted them from observing global features of the distributions. As noted in an earlier activity of the SC, *A* and *D* were attentive to the prominence of "local deviations" in data and this kept them from dealing more freely with global views of data. It is interesting that they did not benefit from this earlier experience. Only after the teacher's intervention they started focusing on relevant information and took into account the variability in the data. Their reasoning about variability evolved then from observing differences between two values, to distinguishing between long and short names, to noticing and informally describing the variability between the groups. They finally arrived (stage 3) at a formulation of a rule or hypothesis that took into account the variability in the data ("*usually, not always*").

In the frequency table task (stage 4), *A* and *D* focused on individual edge values, not noticing the global features of the distribution and ignoring the center interval of the distributions (5 to 7 letters). Possible sources of their difficulties could have been their being novices in the new area of EDA, and the type of representation used, two single frequency tables, which seems complex to analyze and less supportive in terms of displaying general trends. Their initial focus on distribution edges is consistent with other studies, for example, Biehler (2001). Novice students tend to focus on the "least" and the "most" while describing the variability between two distributions using box plots.

The students' insignificant and monotonous use of statistical measures (stage 5) to compare the groups ("*Everything is smaller*") resembles students' reluctance to use averages meaningfully to compare two groups in other studies. There are a number of studies in which students who appeared to use averages to describe a single group or knew how to compute means did not use them to compare two groups (e.g., Bright & Friel, 1998; Watson & Moritz, 1999). Konold et al. (1997) argue that students' reluctance to use averages to compare two groups suggests that they have not developed a sense of average as a measure of a group characteristic, which can be used to represent the group (see also Mokros & Russell, 1995). In addition, students in this study may be seeing averages as only representing middles and having nothing to do with variation.

Throughout their dealing with and comparing the outliers between the groups (stage 6), the students presented a simplistic view of the distributions in order to handle the variability in the data. In their model, resembling a skewed distribution, the majority of the distribution concentrates in one interval, while the less frequent values, the outliers, are positioned in another interval. This model helped them to compare the distributions as following opposing patterns. In their view, the selection of outliers was based on low frequencies, meaning they are exceptional, since they are rare. In that respect, the students' consistent use of "high" and "low" to describe the "long" and "short" surnames in all the dialogues can be attributed to their focus on the variability in frequencies and not only to a careless language flow.

They finally struggled (stage 7) with reading and interpreting the graph they generated (double bar chart, Figure 2). They first practiced their reading of the graph, trying 'vertical' (density) and 'horizontal' (variation in values) interpretations of the variability presented in it. Then they used different local methods to describe the variability in the data. Information they gained in handling the frequency table task helped them in developing a dichotomous model to compare the groups.

The students' development of reasoning about variability in comparing the groups was accompanied by somewhat parallel development of global perception of a distribution as an entity that has typical characteristics such as shape, center, and spread. This perception seems to be a precondition to being able to describe the two distributions as generally similar in shape and variability, but horizontally shifted (USA distribution shifted to the right of the Israeli distribution).

Similar difficulties were demonstrated by eight-grade students working on "prediction" questions about comparing groups (Bakker & Gravemeijer, 2004). These students did not shift a whole shape of a distribution, but reasoned about just the individual bars or the majority (see also Biehler, 2001).

4.2. SUPPORTING FACTORS

The study describes the difficulties and successes of what *A* and *D* did and how they reasoned about variation in the presence of supporting factors that are part of the learning environment in many classes: carefully-planned curricular materials, computer tools, peer collaboration and teacher interventions. It is difficult to tease out, however, what was "naturalistic" about students' actions, and what was an outgrowth of these external factors of the learning environment. What students can and cannot do or think regarding variation is not merely a series of simple natural steps, but rather reactions to and struggles with the challenges and tools (including computer tools, two frequency displays, bar graphs, etc.) that were presented to them at each successive stage of an EDA journey. In particular, students' statistical reasoning and actions were developed throughout by introduction to new cognitive tools and statistical concepts in a supportive learning environment.

Several factors appear to have helped the students develop their statistical reasoning about variability:

a) Students repeatedly experimented with using different informal tools and methods, mostly local in nature (e.g., comparing heights of adjacent bars in a graph) or invented simple models (e.g., dividing the distributions to two subgroups) that partially capture the variability in the data within and between the groups.

b) Students were helped by previous experiences with these data and other sets of data. For example, the dichotomous interpretation of the graph (stage 6) outgrows of previous handling of the statistical measures table.

c) The context of the *Surnames* problem (e.g., the difference between Hebrew and English names) supported *A*'s and *D*'s reasoning in the statistical sphere and provided reasonable explanations to the patterns they observed in the variation. Integration of statistical knowledge and contextual knowledge is considered a fundamental element of statistical thinking (Pfannkuch & Wild, 2004).

d) The incorporation of technological tools enabled students to simply and directly explore data in different forms and experiment with and alter views or displays of data.

e) The interactions with the teacher helped students to adopt a statistical perspective but did not instruct them in exactly what to do. A detailed description of the teacher's role is provided in the following section.

4.3. APPROPRIATION: A LEARNING PROCESS THAT PROMOTES UNDERSTANDING

The data show that most of the learning took place through dialogues between the students themselves but also after brief conversations with the teacher. Of special interest were the teacher's interventions at the students' request (additional examples of such interventions are described in Ben-Zvi & Arcavi, 2001; Ben-Zvi, 2004). These interventions, though short and not necessarily directive, had catalytic effects. They can be characterized in general as "negotiations of meanings" (in the sense of Yackel & Cobb, 1996). More specifically, they are interesting instances of appropriation as a nonsymmetrical, two-way process (in the sense of Moschkovich, 1989). This process takes place, in the zone of proximal development (Vygotsky, 1978, p. 86), when individuals (expert and novices, or teacher and students) engage in a joint activity, each with their own understanding of the task. Students take actions that are shaped by their understanding; the teacher "appropriates" those actions, into her own framework, and provides feedback in the form of her understandings, views of relevance, and pedagogical agenda. Through the teacher's feedback, the students start to review their actions and create new understandings for what they do.

In this study, the teacher appropriated students' utterances with several objectives: to reinforce the legitimacy of an interpretation as the right 'kind' in spite of not being fully correct, to simply refocus

attention on the question, to redirect their attention, to encourage certain initiatives, and implicitly to discourage others (by not referring to certain remarks). The students appropriate from the teacher a reinterpretation of the meaning of what they do. For example, they appropriate from her answers to their inquiries (e.g., what phrasing an hypothesis or interesting phenomena may mean), from her unexpected reactions to their request for explanation (e.g., "*You suggest that there are very short names and very long ones.*"), and from inferring purpose from the teacher's answers to their questions (e.g., "*About the length of surnames. OK?*"). Appropriation by the teacher (to support learning) or by the students (to change the sense they make of what they do) seems to be a central mechanism of enculturation: entering and picking up the points of view of a community or culture (Schoenfeld, 1992; Resnick, 1988). In this process, the teacher is considered as an 'enculturator'. As shown in this study, this mechanism is especially salient when students learn the dispositions that accompany using the subject matter (data analysis) rather than its skills and procedures.

4.4. LIMITATIONS OF THE STUDY

The two students described in this study were considered by their teacher to be both able and verbal. Their choice was aimed to enable the collection and analysis of focused and remarkably detailed data in order to draw, in very fine strokes, the "picture" of their emerging statistical reasoning about variability. Even when a phenomenon seems important and the data interpretation was validated and agreed upon, the question of the idiosyncrasy of the identified phenomenon may remain open. Therefore, in other studies, the data and interpretations from students in the same class or from other classes assist in checking for generalizability of the phenomena (cf., Ben-Zvi, 2002).

In presenting the students with tasks based on comparing two groups of equal size, some complications are avoided. This is both an advantage and disadvantage for the overall aims of this study. Research shows that the group comparison problem is one that students do not initially know how to approach and the challenge may remain even after extended periods of instruction (e.g., Bakker & Gravemeijer, 2004). Avoiding some of the complexity of proportional reasoning, the key for handling groups of different size, simplifies the task and may help researchers focus on and expose students' reasoning about variability. In this study, students were "pushed" to consider other complex statistical issues, such as integrating measures of variation and center and comparing measures *within* each group and *between* groups. However, it should be acknowledged that the study of students' statistical reasoning about variability in comparing groups is not complete without incorporating tasks of comparing unequal data sets.

5. IMPLICATIONS

The idiosyncratic aspects of this study restrict the provision of broad recommendations. However, several conclusions that are tied to specifics of this study and its results, in the context of results from similar studies, can be drawn. The learning processes described in this paper took place in a carefully designed environment. This environment included: a curriculum built on the basis of expert views of EDA as a sequence of semi-structured, yet open, leading questions within the context of extended meaningful problem situations (Ben-Zvi & Arcavi, 1998), timely and non-directive interventions by the teacher as representative of the discipline in the classroom (cf., Voigt, 1995), and computerized tools that enable students to handle complex actions (change of representations, scaling, deletions, restructuring of tables, etc.) without having to engage in too much technical work, leaving time and energy for conceptual discussions (cf., Ben-Zvi, 2000).

In learning environments of this kind, from the very beginning students encounter, develop, and work with ideas, concepts, cognitive tools and dispositions related to the culture of EDA, such as making hypotheses, summarizing data, recognizing trends and variability, identifying interesting phenomena, comparing distributions and handling numerical, tabular and graphical data representations. Skills, procedures and strategies, such as creating and interpreting graphs and tables or calculating statistical measures, are learned as integrated in the context and at the service of the main ideas of EDA.

It can be expected that beginning students will have difficulties, of the type described, when confronting the problem situations of the curriculum. However, it is proposed that what *A* and *D* experienced should be an integral and inevitable component of a meaningful learning process if it is to have lasting effects. If students were to work in environments such as the above, teachers are likely to encounter the following learning phenomena:

* Students' prior knowledge would and should be engaged in interesting and surprising ways, possibly hindering progress in some instances but making the basis for construction of new knowledge in others,

* many questions that would either make little sense to the students, or, alternatively, will be re-interpreted and answered in different ways than intended, and

* students' work that would inevitably be based on partial understandings, which will grow and evolve.

This study suggests that in order to help students gradually build a sense of the meaning of the data and statistical task with which they engage, multiple factors can and should be planned. These include appropriate teacher guidance, peer work and interactions, and more importantly, ongoing cycles of experiences with realistic problem situations.

Given that it is difficult to tease out the effects of what students learned or could or couldn't do from the enculturation processes and support of the teacher, further study is recommended that focus more attention on the role of teachers and what they should do, or learn to do, in order to promote statistical reasoning about variability. Much of students' progress in the current study is influenced by their interactions with the teacher that helped them adopt the statistical perspective but did not instruct them in exactly what to do or how to reason. The role of the teacher which is considered as an 'enculturator' deserves further exploration.

It is generally recommended that students be provided with multiple opportunities to engage with data in group-comparison tasks. The role of comparing unequal-size groups in promoting reasoning about variability, which was not studied here, should be further explored. The students in this study have gained from reading and interpreting multiple types of conventional data representations. The role of student-invented data representations and new graphical tools available through educational software and Internet has to be investigated to better expose the many ways variability is noticed, measured, and modeled by students. It is hoped that the complexity involved in group-comparison tasks can push students to think about the meaning of what they do and how they reason in statistics, develop relevant actions and interpretations, and be more critical of their actions and interpretations.

REFERENCES

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

Ben-Zvi, D. (2000). Toward understanding of the role of technological tools in statistical learning. *Mathematical Thinking and Learning, 2*(1&2), 127–155.

Ben-Zvi, D. (2002). Seventh grade students' sense making of data and data representations. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics, Cape Town, South Africa*. [CD-ROM] Voorburg, The Netherlands: International Statistical Institute.

Ben-Zvi, D. (2004). Reasoning about Data Analysis. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 121–146). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Ben-Zvi, D., & Arcavi, A. (1998). Toward a characterization and understanding of students' learning in an interactive statistics environment. In L. Pereira-Mendoza (Ed.), *Proceedings of the Fifth International Conference on Teaching Statistics* (Vol. 2, pp. 647–653). Voorburg, The Netherlands: International Statistical Institute.

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

Ben-Zvi, D., & Friedlander, A. (1997a). *Statistical investigations with spreadsheets—Student's workbook* (In Hebrew). Rehovot, Israel: Weizmann Institute of Science.

Ben-Zvi, D., & Friedlander, A. (1997b). Statistical thinking in a technological environment. In J. B. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics* (pp. 45–55). Voorburg, The Netherlands: International Statistical Institute.

Ben-Zvi, D., & Ozruso, G. (2001). *Statistical investigations with spreadsheets—Teacher's guide* (In Hebrew). Rehovot, Israel: Weizmann Institute of Science.

Biehler, R. (1993). Software tools and mathematics education: The case of statistics. In C. Keitel & K. Ruthven (Eds.), *Learning from computers: Mathematics education and technology* (pp. 68–100). Berlin: Springer-Verlag.

Biehler, R. (1997). Software for learning and for doing statistics. *International Statistical Review*, *65*(2), 167–189.

Biehler, R. (2001, August). Developing and assessing students' reasoning in comparing statistical distributions in computer supported statistics courses. Paper presented at the Second International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-2), Armidale, Australia.

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

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

Gal, I., & Garfield, J. B. (Eds.). (1997). *The assessment challenge in statistics education*. Amsterdam, Netherlands: IOS Press.

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

Garfield, J. (1995). How students learn statistics. *International Statistical Review 63*(1), 25–34.

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

Hershkowitz, R. (1999). Where in shared knowledge is the individual knowledge hidden? In O. Zaslavsky (Ed.) *Proceedings of the 23rd Conference of the International Group for the Psychology of Mathematics Education*, *I*, (pp. 9–24). Haifa, Israel: The Technion.

Hershkowitz, R., Dreyfus, T., Schwarz, B., Ben-Zvi, D., Friedlander, A., Hadas, N., Resnick, T., & Tabach, M. (2002). Mathematics curriculum development for computerized environments: A designer-researcher-teacher-learner activity. In L. D. English (Ed.), *Handbook of international research in mathematics education* (pp. 657–694). London: Erlbaum.

Hunt, D. N. (1995). Teaching statistical concepts using spreadsheets. In the *Proceedings of the 1995 Conference of the Association of Statistics Lecturers in Universities*. Nottingham, UK: The Teaching Statistics Trust. [Online: http://www.mis.coventry.ac.uk/~nhunt/aslu.htm]

Konold, C (2002). Teaching concepts rather than conventions. *New England Journal of Mathematics*, *34*(2), 69–81.

Konold, C., & Higgins, T. (2003). Reasoning about data. In J. Kilpatrick, W. G. Martin & D. E. Schifter (Eds.), *A research companion to principles and standards for school mathematics*, (pp. 193–215). Reston, VA: National Council of Teachers of Mathematics.

Konold, C., Pollatsek, A., Well, A., & Gagnon, A. (1997). Students analyzing data: Research of critical barriers. In J. B. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics*, (pp. 151–167). Voorburg, The Netherlands: International Statistical Institute.

Lampert. M. (1990). When the problem is not the question and the solution is not the answer: Mathematical knowing and teaching. *American Educational Research Journal, 27*, 29–63.

Magidson, S. (1992, April). From the laboratory to the classroom: A technology-intensive curriculum for functions and graphs. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Makar, K., & Confrey, J. (2004). Secondary teachers' statistical reasoning in comparing two groups. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 353–374). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Meira, L. R. (1991*)*. Explorations of mathematical sense-making: An activity-oriented view of children's use and design of material displays. An unpublished Ph.D. dissertation, Berkeley, CA: University of California.

Meletiou, M. (2002). Conceptions of variation: A literature review. *Statistics Education Research Journal, 1*(1), 46–52.

Meletiou, M., & Lee, C. (2002). Student understanding of histograms: A stumbling stone to the development of intuitions about variation. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics, Cape Town, South Africa.* [CDROM] Voorburg, The Netherlands: International Statistical Institute.

Mokros, J., & Russell, S. J. (1995). Children's Concepts of Average and Representativeness. *Journal for Research in Mathematics Education, 26*(1), 20–39.

Moschkovich, J. D. (1989). Constructing a problem space through appropriation: A case study of guided computer exploration of linear functions. An unpublished manuscript available from the author.

Moschkovich, J. D., Schoenfeld, A. H., & Arcavi, A. A. (1993). Aspects of understanding: On multiple perspectives and representations of linear relations, and connections among them. In T. Romberg, E. Fennema & T. Carpenter (Eds.), *Integrating Research on the Graphical Representation of Function*, (pp. 69–100). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

Resnick, L. (1988). Treating mathematics as an ill-structured discipline. In R. Charles & E. Silver (Eds.), *The teaching and assessing of mathematical problem solving* (pp. 32–60). Reston, VA: National Council of Teachers of Mathematics.

Resnick, T., & Tabach, M. (1999). *Touring the land of Oz - algebra with computers for Grade Seven* (in Hebrew). Rehovot, Israel: Weizmann Institute of Science.

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 334–370). New York: Macmillan.

Schoenfeld, A. H. (1994). Some notes on the enterprise (research in collegiate mathematics education, that is). *Conference Board of the Mathematical Sciences Issues in Mathematics Education, 4*, 1–19.

Shaughnessy, J. M., & Ciancetta, M. (2002). Students' understanding of variability in a probability environment. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics*, *Cape Town, South Africa.* [CDROM] Voorburg, The Netherlands: International Statistical Institute.

Shaughnessy, J. M., Garfield, J., & Greer, B. (1996). Data handling. In A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics education* (Vol. I, pp. 205–237). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*, (Edited by M. Cole, V. John-Steiner, S. Scribner, & E. Souberman). Cambridge, MA: Harvard University Press.

Voigt, J. (1995). Thematic patterns of interaction and sociomathematical norms. In P. Cobb & H. Bauersfeld (Eds.), *Emergence of mathematical meaning: Interaction in classroom cultures*, (pp. 163–201). Hillsdale, NJ: Erlbaum.

Watson, J. M. (2001). Longitudinal development of inferential reasoning by school students. *Educational Studies in Mathematics, 47*, 337–372.

Watson, J. M., & Kelly, B. A. (2002). Can grade 3 students learn about variation? In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics*, *Cape Town, South Africa*. [CDROM] Voorburg, The Netherlands: International Statistical Institute.

Watson, J. M., Kelly, B. A., Callingham, R. A., & Shaughnessy, J. M. (2003). The measurement of school students' understanding of statistical variation. *International Journal of Mathematical Education in Science and Technology*, *34*(1), 1–29.

Watson, J. M., & Moritz, J. B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics, 37*(2), 145–168.

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

Yackel, E., & Cobb, P. (1996). Socio-mathematical norms, argumentation and autonomy in mathematics. *Journal for Research in Mathematics Education, 27*(4), 458–477.

SOFTWARE

Excel, Microsoft Corporation, http://www.microsoft.com/office/excel/.

Fathom (Fathom Dynamic Statistics Software), B. Finzer, Key Curriculum Press, 1150 65th Street, Emeryville, CA 94608, USA. http://www.keypress.com/fathom/.

Mini-Tools, Peabody College, Vanderbilt University, principal investigator: P. Cobb, http://peabody.vanderbilt.edu/depts/tandl/mted/Proj6_CMT/6MiniTools.html.

Tinkerplots, the Statistics Education Research Group at the University of Massachusetts, Amherst, principal investigator: C. Konold, http://www.umass.edu/srri/serg/projects/tp/tpmain.html.

DANI BEN-ZVI
Faculty of Education
University of Haifa
Mount Carmel
Haifa 31905
Israel

REASONING ABOUT SHAPE AS A PATTERN IN VARIABILITY

ARTHUR BAKKER
Freudenthal Institute
Utrecht University
The Netherlands
a.bakker@ioe.ac.uk

SUMMARY

This paper examines ways in which coherent reasoning about key concepts such as variability, sampling, data, and distribution can be developed as part of statistics education. Instructional activities that could support such reasoning were developed through design research conducted with students in grades 7 and 8. Results are reported from a teaching experiment with grade 8 students that employed two instructional activities in order to learn more about their conceptual development. A "growing a sample" activity had students think about what happens to the graph when bigger samples are taken, followed by an activity requiring reasoning about shape of data. The results suggest that the instructional activities enable conceptual growth. Last, implications for teaching, assessment and research are discussed.

Keywords: Design research; Distribution; Instructional activities; Middle school level; Sampling

1. BACKGROUND OF THE RESEARCH

The first time I visited an American classroom I attended a statistics lesson in grade 5. When the teacher asked a question that sounded statistical but did not require a measure of center, one student, Malcolm, thoughtlessly muttered "meanmedianmode," as if it were one word. My impression was that these students had been drilled to calculate mean, median, and mode, and to draw bar graphs, but did not use their common sense in answering statistical questions. This small incident exemplifies what a litany of research in statistics education reports on: too often, students learn statistics as a set of techniques that they do not apply sensibly. Even if they have learned to calculate mean, median, mode, and to draw histograms and box plots, they mostly do not understand that they can use a mean as a group descriptor when comparing two data sets—to give one example that is well documented (Konold & Higgins, 2003; McGatha, Cobb, & McClain, 2002; Mokros & Russell, 1995). This problem is not typically American; it also applies to the Dutch context, but to a lesser extent. The reason for this is probably that Dutch students mostly learn statistical concepts and graphs such as median, mode, histogram, and box plot about three years later than in the USA.

Despite differences between the curricula in different countries, the underlying problem remains the same: students generally lack the necessary conceptual understanding for analyzing data with the statistical techniques they have learned. The problem many statistics educators encounter is that students tend to perceive data just as a series of individual cases (a case-oriented view), and not as a whole that has characteristics that are not visible in any of the individual cases (an aggregate view). Hancock, Kaput, and Goldsmith (1992) note that students need to mentally construct such an aggregate before they can perceive a data set as a whole. Many researchers have encountered the same problem and experienced its persistency (e.g. Ben-Zvi & Arcavi, 2001; Wilensky, 1997).

The above implies that students need to develop a conceptual structure with which they can conceive data sets as aggregates. Konold and Pollatsek (2002) argue that students need to develop a conceptual understanding of signal and noise in order to understand what an average value is about in relation to the variation around that value. The present paper distinguishes two types of signals in noisy processes or patterns in variability. First, the signal can be a true value with error as noise around it. Such signals are apparent in repeated measurements of one item (Petrosino, Lehrer, & Schauble, 2003). The "center clump" is then an indication of where the true value probably is. Second, the signal might be a distribution, such as the shape of a smooth bell curve of the normal distribution, with which we model data. The noise in that case is the variation around that smooth curve. In either type of pattern, it is evident that students need good conceptual understanding before they can recognize signals in noisy processes. This paper focuses on the second type of signal in noisy processes or, other words, *shape as a pattern in variability*.

The concept of distribution is a structure with which students can conceive aggregate features of data sets (Cobb, 1999; Gravemeijer, 1999). Petrosino et al. (2003, p. 132) write: "Distribution could afford an organizing conceptual structure for thinking about variability located within a more general context of data modeling." Of course, distribution is a very advanced concept that, in its full complexity, is far beyond the scope of middle school students (Wilensky, 1997). Nevertheless, it is possible to address the issue of how data are distributed from an informal situational level onwards by focusing on shape (Bakker & Gravemeijer, 2004; Cobb, 1999; Russell & Corwin, 1989). In the research of Cobb, McClain, and Gravemeijer (2003), students came to reason with "hills" to indicate "majorities" of data sets, which are informal terms to describe the hill shape of data sets and areas in graphs where most data points seemed to be. Bakker and Gravemeijer (2004) report that students reasoned how a "bump" would shift if older students were measured, and what would happen with the bump if the sample "grew", i.e., began to include more and more cases. The seventh-grade students in their study came to see a pattern in the variability of different phenomena such as weight, height, and wingspan of birds.

In the first teaching experiment I conducted in grade 7 (Bakker, 2004), I focused on the concept of distribution, but this turned out to be too limited. Sampling, for instance, is also crucial to address in an early stage of statistical data analysis (see also Bakker & Gravemeijer, 2004). The research presented here focused on a broader set of key concepts that students, in my view, need to develop in order to analyze data in a meaningful way: variability, sampling, data, and distribution (cf. Garfield & Ben-Zvi, 2004). The main question of the overall research was, how can we promote coherent reasoning about variability, sampling, data, and distribution in a way that is meaningful for students with little statistical background?

The learning process aimed at in this research can be characterized as "guided reinvention" (Freudenthal, 1991). Students were stimulated to contribute their own ideas, strategies, and language in solving statistical problems (reinvention), but they were also provided with increasingly sophisticated ways to describe how data were distributed and to characterize data sets (planned guidance).

In this paper I report results from a teaching experiment that employed two instructional activities I developed. The paper analyzes students' learning process in order to learn more about their development of key concepts underlying statistical data analysis, especially variability, sampling, data, and distribution. The two activities used seemed particularly promising for fostering coherent reasoning about these key concepts and were developed using a cyclic approach of designing instructional materials, testing them during classroom-based teaching experiments, analyzing students' learning process, and revising the instructional materials. The first instructional activity, growing a sample or a data set, is an elaboration of an activity described in Bakker and Gravemeijer (2004). The second activity involved reasoning about shapes that students themselves had proposed. In grade 7, it turned out to be difficult for students to reason with shape, except for high achievers who reasoned with bumps. The teaching experiment reported here was therefore carried out in a higher grade (8th).

Below, I first elaborate on the methodological approach of design research employed in this eighth-grade teaching experiment, and then describe the subjects, data collection, and method of

analysis. Results are then presented regarding students' reasoning during two instructional activities, "growing a sample" and "reasoning about shapes". Finally, the results and their limitations are discussed, as well as implications for teaching, assessment, and research.

2. METHODOLOGY AND SUBJECTS

2.1. DESIGN RESEARCH

To answer the question of how coherent reasoning about variability, sampling, data, and distribution could be promoted, I needed to design instructional activities that could support such reasoning as well as to understand *how* those activities supported students' conceptual development. If the kind of education aimed at is not yet available, the required conditions first need to be created. Instructional design is therefore an important part of the research presented here. In general, if you want to change something you have to understand it and if you want to understand something you have to change it. In this approach, design and "research" are highly intertwined, and it will not surprise the reader that this type of research falls under the general heading of "design research." Design research typically involves cycles of three phases: a preparation and a design phase (of instructional materials for example), teaching experiments, and retrospective analyses (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003; Gravemeijer, 1994).

1. Preparation and design phase. In the research presented here, the preparation phase consisted of a literature survey, a historical study of the statistical concepts and graphs at issue (Bakker, 2003), and the first reformulation of a Hypothetical Learning Trajectory (HLT), which Simon (1995) defines as follows: "A hypothetical learning trajectory is made up of three components: the learning goal that defines the direction, the learning activities, and the hypothetical learning process – a prediction of how the students' thinking and understanding will evolve in the context of the learning activities" (p. 136). The hypothetical learning trajectory of the present study was to support students in reasoning about aspects of distribution and sampling using increasingly sophisticated concepts and graphs. Further details about hypothetical learning trajectories can be found in a special issue of *Mathematical Thinking and Learning* devoted to this topic: 6(2).

2. Teaching experiment. The HLT is tested and possibly revised during a teaching experiment. The anticipations formulated in the HLT give guidance to both teacher and researcher of what to focus on during instruction, interviewing, and observation. The teacher and researcher can adjust their original plans if new ideas seem to be better, so they need not wait till the end of the teaching experiment to change activities or even the end goal.

3. Retrospective analysis. The retrospective analysis is meant to find out if the anticipations of the HLT were right, to find patterns in students' learning processes and to understand the role of the instructional materials (activities, software). New insights mostly lead to the revision of the instructional materials, the end goals, or the route to be taken next time and a revised HLT that can guide the next teaching experiment.

Overall, the idea behind developing an HLT is not to design the perfect instructional sequence, which in my view does not exist, but to provide empirically grounded results that others can adjust to their local circumstances. The HLT remains hypothetical because each situation, each teacher, and each class is different. Yet patterns can be found in students' learning that are similar across different teaching experiments (Bakker, 2004). Those patterns and the insights of how particular instructional activities support students in particular kinds of reasoning can be the basis for a more general instructional theory of how a particular domain can be taught.

2.2. SETUP, SUBJECTS, DATA COLLECTION, AND ANALYSIS

Setup. This paper focuses on the fourth and the sixth lessons of a series of ten lessons, each 50 minutes long. In these specific lessons, on which the restrospective analyses also centered, students reasoned about larger and larger samples and about the shape of distributions. Half of the lessons

were carried out in a computer lab and as part of them students used two software minitools (Cobb, Gravemeijer, Bowers, & Doorman, 1997), simple Java applets with which they analyzed data sets on, for instance, battery life span, car colours, and salaries (Figure 1). As a researcher I was responsible for the instructional materials and the teacher was responsible for the teaching, though we discussed in advance on a weekly basis both the materials and appropriate teaching style. Three pre-service teachers served as assistants and helped with videotaping and interviewing students and with analyzing the data.

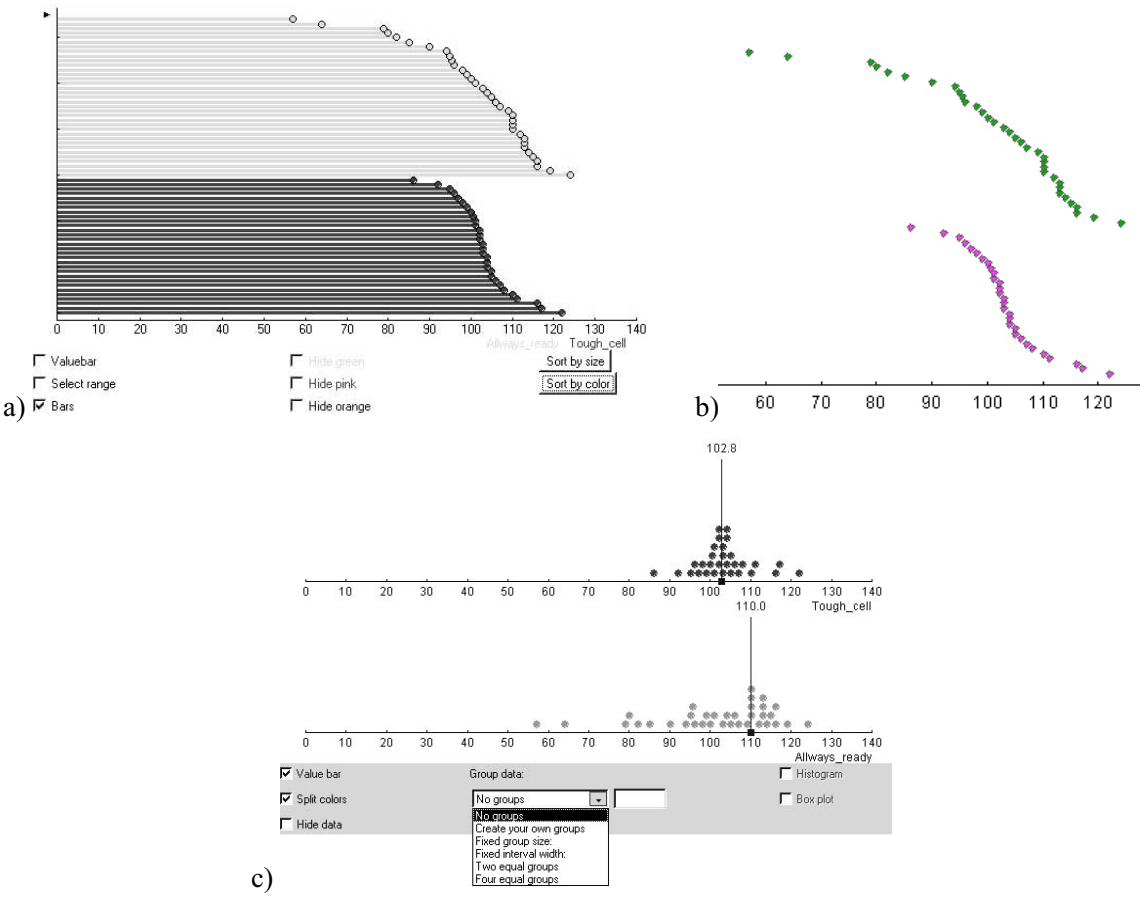

Figure 1. a) Minitool 1 showing a value-bar graph of battery life spans in hours of two brands. b) Minitool 1, but with bars hidden. c) Minitool 2 showing a dot plot of the same data sets.

Subjects. The teaching experiment was carried out in an eighth-grade class with 30 students in a public school in the center of the Dutch city of Utrecht in the fall of 2001. The students in this study were being prepared for pre-university (*vwo*) or higher vocational education (*havo*). The top 35-40% of Dutch students attend these types of education. The remaining 60-65% of students are prepared for other types of vocational education (*vmbo*). Other relevant background information is that school textbooks play a central role in the practice of Dutch mathematics education. Students are expected to be able to work through the tasks by themselves, with the teacher available to help them if necessary. As a consequence, tasks are broken down into very small steps and real problem solving is rare. Students' answers tend to be superficial, in part because they have to deal with about eight different problem contexts per lesson (Van den Boer, 2003). The students in the class reported on here were not used to whole-class discussions, but rather to be "taken by the hand" as the teacher called it; they were characterized by the three assistants as "passive but willing to coorporate." These eighth-grade students had no prior instruction in statistics; they were acquainted with bar and line graphs, but not with dot plots, histograms, or box plots. Students already knew the mean from calculating their report grades, but mode and median were not introduced until the second half of the instructional sequence after variability, data, sampling, and shape had been topics of discussion.

Data collection. The collected data on which the results presented in this paper are based (regarding the fourth and sixth lessons) include student work, field notes, and the audio and video recordings of class activities that the three assistants and I made in the classroom. An essential part of the data corpus was the set of mini-interviews we held during the lessons; they varied from about twenty seconds to four minutes, and were meant to find out what concepts and graphs meant for students, or how the mini-tools were used. These mini-interviews influenced students' learning because they often stimulated reflection. However, I think that the validity of the research was not put in danger by this, since the aim was to find out how students learned to reason with shape or distribution, not whether teaching the sequence in other eighth-grade classes would lead to the same results in the same number of lessons. Furthermore, the interview questions were planned in advance, and discussed with the assistants.

Analysis. For the analysis, I read the transcripts, watched the videotapes, and formulated conjectures on students' learning on the basis of episodes identified in the transcripts and video. The generated conjectures were tested against other episodes and the rest of the collected data (student work, field observations, and tests) in the next round of analysis (triangulation). Then the whole generating and testing process was repeated, a method resembling Glaser and Strauss's constant comparative method (Glaser & Strauss, 1967). About one quarter of the episodes, including those discussed in this paper, and the conjectures belonging to these episodes were judged by the three assistants who attended the teaching experiment. The amount of agreement among judges was very high, over 95%. Only the conjectures that all of us agreed upon were kept. An example of a conjecture that was confirmed was that students tended to group data sets, real or imagined, into three groups of low, "average", and high values.

For analyzing students' reasoning with diagrams I used the semiotics of Peirce (1976), in particular his concepts of diagrammatic reasoning and hypostatic abstraction. *Diagrammatic reasoning* involves three steps: constructing a diagram, experimenting with it, and reflecting upon the results. An important part of the reflection step is to describe what is seen in diagrams (bars, dots, relationships, shapes). The process of describing qualities of those objects can be called *predication*. Van Oers (2000) uses the following definition: "Predication is the process of attaching extra quality to an object of common attention (such as a situation, topic or theme) and, by doing so, making it distinct from others" (p. 150). Next, *hypostatic abstraction* is one of the forms of abstraction that Peirce distinguished: a predicate becomes an object in itself that can have characteristics. This is linguistically reflected in the transition from a predicate (e.g. most, lying out) to a noun (majority, outlier). Note that this paper focuses only on predication or hypostatic abstraction in instances of diagrammatic reasoning.

3. RESULTS

This section presents the analysis of students' reasoning during the two instructional activities on which this paper is focused. The first, carried out in the fourth lesson, is reasoning about larger and larger samples, or larger and larger data sets. I call this activity "growing a sample" (following Konold & Pollatsek, 2002), though some readers may prefer to call it "growing a data set." The term "sample" is preferred here because one intention of the activity was to let students think about samples versus populations. The second activity, carried out in the sixth lesson, is reasoning about shapes of these data sets. For each activity, I first summarize the hypothetical learning trajectory (HLT) of that lesson, and then present the analysis.

3.1. GROWING A SAMPLE

The overall goal of the growing samples activity as formulated in the hypothetical learning trajectory for this fourth lesson was to let students reason about shape in relation to sampling and distribution aspects in the context of weight. The idea was to start with students' own ideas and guide them toward more conventional notions and representations.

Before getting to the growing samples activity, in a previous (third) lesson students had answered the question of how many eighth graders could go into a hot air balloon if normally eight adults (apart from the balloon pilot) were allowed. They also had made a prediction of a weight graph of eighth graders, without having any data available yet. It was apparent that students tended to choose small group sizes such as 10 or 15. The students had also had two lessons of experience with two the minitools. The first minitool supplies a value-bar graph in which each bar has a length corresponding to the data value it represents (Figure 1a); the second minitool provides a dot plot (Figure 1c). In the first minitool, students can organize data, for instance by sorting or hiding subsets of data, and by sorting the data by size. They can also hide the bars, so that they only see the endpoints of the bars (Figure 1b). In the second minitool, these endpoints have been collapsed onto the axis. Students can organize data with different options, for instance making their own groups, two equal groups (precursor to the median), four equal groups (precursor to box plot), and fixed interval width (precursor to histogram).

The activity of growing a sample built on the balloon activity. It consisted of three cycles of making sketches of a hypothetical situation and comparing those sketches with graphs displaying real data sets. In the first cycle students had to make a graph of their own choice of a predicted weight data set with sample size 10. The results were discussed by the teacher to challenge this small sample size, and in the subsequent cycles students had to predict larger data sets (one class, three classes, all students in the province). Three such cycles took place as described below. The teacher and I tried to strike a balance between engaging students in statistical reasoning and allowing their own terminology on the one hand, and guiding them in using conventional and more precise notions and graphical representations on the other.

First cycle of growing a sample

The text of the student activity sheet for the fourth lesson started as follows:

Last week you made graphs of predicted data for a balloon pilot. During this lesson you will get to see real weight data of students from another school. We are going to investigate the influence of the sample size on the shape of the graph.

a. Predict a graph of ten data values, for example with the dots of minitool 2.

The sample size of ten was chosen because the students had found that size reasonable after the first lesson in the context of testing the life span of batteries. Figure 2a shows examples for three different types of diagrams the students made to show their predictions: there were three value-bar graphs (such as in minitool 1, e.g., Ruud's diagram), eight with only the endpoints (such as with the option of minitool 1 to "hide bars", e.g., Chris's diagram) and the remaining nineteen plots were dot plots (such as in minitool 2, e.g., Sandra's diagram). This means that their graphs were heavily influenced by their experiences with the minitools.

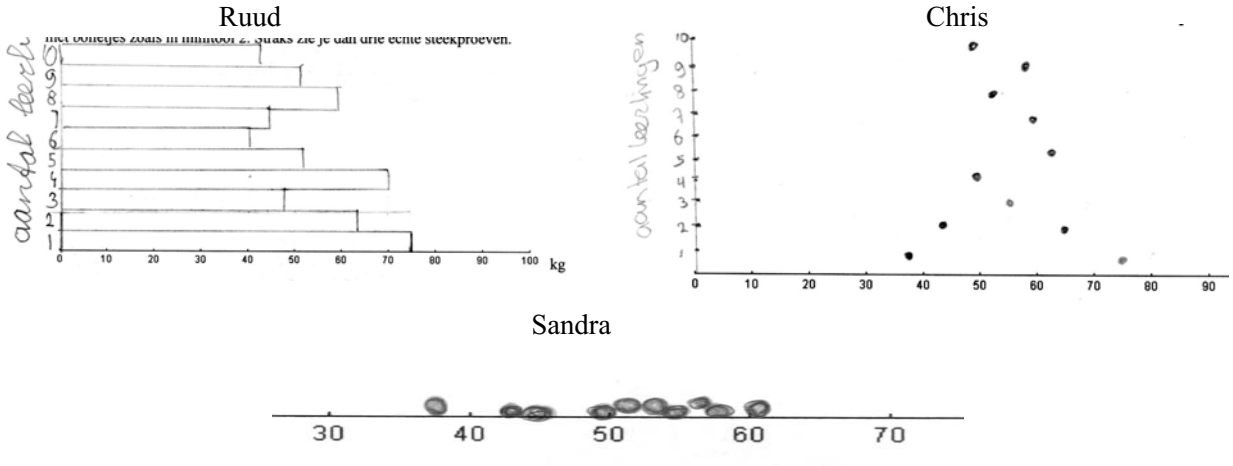

Figure 2a. Student predictions (Ruud, Chris, and Sandra) for ten data points (weight in kg)

For the remainder of this section, the figures and written explanations of these three students are demonstrated, because their work gives an impression of the variety of the whole class. The learning abilities of these students varied considerably: Ruud and Chris's report grades were in the bottom third of the class whereas Sandra had the best overall report score of the class across all subjects. I have chosen those three students because their diagrams represent all types of diagrams made in this class, also for other cycles of growing a sample.

To stimulate the reflection on the graphs, the teacher showed three samples of ten data points on the blackboard and students had to compare their own graphs (Figure 2a) with the graphs of the real data sets (Figure 2b).

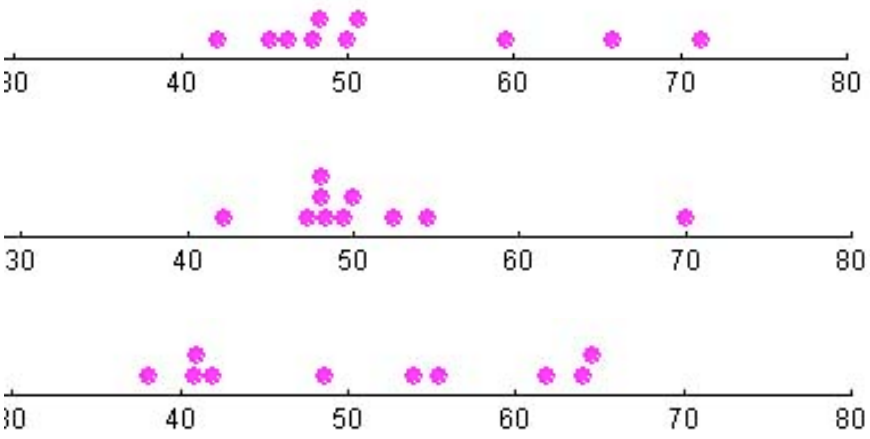

Figure 2b. Three real data sets in minitool 2

b. You get to see three different samples of size 10. Are they different than your own prediction? Describe the differences.

The reason for showing three small samples was to show the variation among these samples. There were no clear indications, though, that students conceived this variation as a sign that the sample size was too small for drawing conclusions, but they generally agreed that larger samples were more reliable. There was a short class discussion about the graphs with real data before students worked for themselves again. The point relevant to the analysis is that students started using predicates to describe aggregate features of the graphs. Please note that a grammatical translation into English of ungrammatical spoken Dutch does not always sound very authentic.

Teacher:	We're going to look at these three different ones [samples in Figure 2b]. Can anyone say something yet? Give it a try.
Jacob:	In the middle [graph], there are more together.
Teacher:	Here [pointing to the middle graph of Figure 2b] there are many more together, clumped or something like that. Who can mention other differences?
Jacob:	Well, uh, the lowest, I think it's all the furthest apart.
Teacher:	Those are all the furthest apart. Here [in the middle graph] they are in one clump. Are there any other things you notice, Gigi?
Gigi:	Yes, the middle one has just one at 70. [This is a case-oriented view.]
Teacher:	There's only one at 70 and the rest are at 60 or lower? Yes? Can you say something about the mean perhaps?
Rick:	The mean is usually somewhere around 50.

The written answers of the three students were the following:

Ruud:	Mine looks very much like what is on the blackboard.
Chris:	The middle-most [diagram on the blackboard] best resembles mine because the weights are close together and that is also the case in my graph. It lies between 35 and 75 [kg].
Sandra:	The other [real data] are more weights together and mine are further apart.

Ruud's answer is not very specific, like most of the written answers in the first cycle of growing samples. Chris used the predicate "close together" and added numbers to indicate the range, probably as an indication of spread. Sandra used such terms as "together" and "further apart," which address spread. The students in the class used common predicates such as "together," "spread out" and "further apart" to describe features of the data set or the graph. For the analysis it is important to note that the students used predicates (together, apart) and no nouns (spread, average) in this first cycle of growing samples. Spread can only become an object-like concept, something that can be talked about and reasoned with, if it is a noun. In the semiotic theory of Peirce (1976), such transitions from the predicate "the dots are spread out" to "the spread is large" are important steps in the formation of concepts.

Second cycle of growing a sample

The students generally understood that larger samples would be more reliable. With the feedback students had received after discussing the samples of ten data points in dot plots, students had to predict the weight graph of a whole class of 27 students and of three classes with 67 students (27 and 67 were the sample sizes of the real data sets of eighth graders of another school).

c. We will now have a look how the graph changes with larger samples. Predict a sample of 27 students (one class) and of 67 students (three classes).

d. You now get to see real samples of those sizes. Describe the differences. You can use words such as majority, outliers, spread, average.

During this second cycle, all of the students made dot plots, probably because the teacher had shown dot plots on the blackboard, and because dot plots are less laborious to draw than value bars (only one student started with a value-bar graph for the sample of 27, but switched to a dot plot for the sample of 67). The hint on statistical terms was added to make sure that students' answers would not be too superficial, as often happened before, and to stimulate them to use such notions in their reasoning. It was also important for the research to know what these terms meant for them. When the teacher showed the two graphs with real data, there was once again a short class discussion in which the teacher capitalized on the question of why most student prediction now looked pretty much like what was on the blackboard, whereas with the earlier predictions there was much more variation. No student had a reasonable explanation, which indicates that this was an advanced question. When comparing their own graphs (Figure 3a) with real data (Figue 3b), the same three students wrote:

Ruud:	My spread is different.
Chris:	Mine resembles the sample, but I have more people around a certain weight and I do not really have outliers, because I have 10 about the 70 and 80 and the real sample has only 6 around the 70 and 80.
Sandra:	With the 27 there are outliers and there is spread; with the 67 there are more together and more around the average.

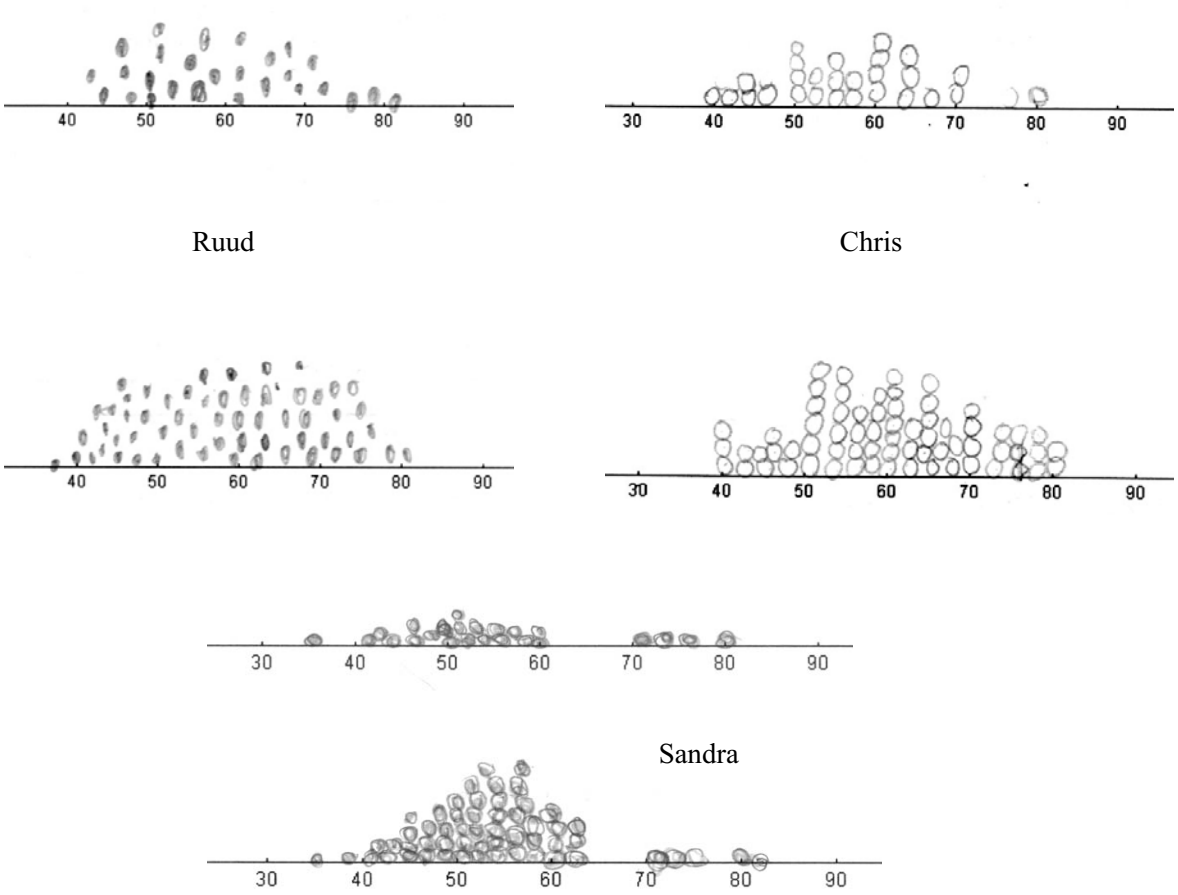

Figure 3a. Predicted graphs for one and for three classes by Ruud, Chris, and Sandra

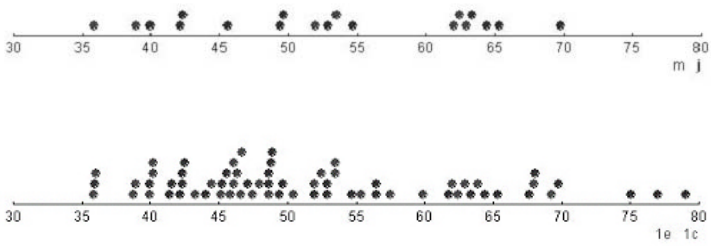

Figure 3b. Real data sets of size 27 and 67 of students from another school

Here, Ruud addressed the issue of spread. Chris was more explicit about a particular area in her graph, the category of high values. She also correctly used the term "sample," which was newly introduced in the second lesson. Sandra used the term "outliers" at this stage, by which students meant "extreme values" (not necessarily exceptional or suspect values). She also seemed to locate the average somewhere and to understand that many students are about average. These examples illustrate

that students used statistical notions for describing properties of the data and diagrams. From a statistical point of view, these terms were not very precise. With "mean" students generally meant "about average" or "the middle typical group"; with "spread" they meant "how far the data lie apart". And with "sample" they seemed to mean just a bunch of people, not necessarily the data as being representative for a population (cf. Schwartz et al., 1998).

In contrast to the first cycle of growing a sample, students used nouns instead of just predicates for comparing the diagrams. Ruud (like others) used the noun "spread," whereas students earlier used only predicates such as "spread out." Of course, this does not always imply that if students use these nouns that they are thinking of the right concept. Statistically, however, it makes a difference whether we say, "the dots are spread out" or "the spread is large." In the latter case, spread is an object-like entity that can have particular aggregate characteristics that can be measured (for instance by the range, the interquartile range, or the standard deviation). Other notions, outliers, sample, and average, are now used as nouns, that is as conceptual objects that can be talked about and reasoned with.

Third cycle of growing a sample

The aim of the hypothetical learning trajectory was that students would come to draw continuous shapes and reason about them using statistical terms. During teaching experiments in the seventh-grade experiments (Bakker & Gravemeijer, 2004), experiments in two American sixth-grade classes, and a visit to an American group of ninth graders, reasoning with continuous shapes turned out to be difficult to accomplish, even if it was asked for. It often seemed impossible to nudge students toward drawing the general, continuous shape of data sets represented in dot plots. At best, students drew spiky lines just above the dots. This underlines that students have to construct something new (a notion of signal, shape, or distribution) with which they can look differently at the data or the variable phenomenon.

In this last cycle of growing the sample, the task was to make a graph showing data of all students in the city, not necessarily with dots. The intention of asking this was to stimulate students to use continuous shapes and dynamically relate samples to populations, without making this distinction between sample and population explicit yet. The conjecture was that this transition from a discrete plurality of data values to a continuous entity of a distribution is important to foster a notion of distribution as an object-like entity with which students could model data and describe aggregate properties of data sets. The task proceeded as follows:

e. Make a weight graph of a sample of all eighth graders in Utrecht. You need not draw dots. It is the shape of the graph that is important.

f. Describe the shape of your graph and explain why you have drawn that shape.

The figure of the same three students are presented in Figure 4 and their written explanations were:

Ruud:	Because the average [values are] roughly between 50 and 60 kg.
Chris:	I think it is a pyramid shape. I have drawn my graph like that because I found it easy to make and easy to read.
Sandra:	Because most are around the average and there are outliers at 30 and 80 [kg].

Ruud's answer focused on the average group, or "modal clump" as Konold and colleagues (2002) call such groups in the center. During an interview after the fourth lesson, Ruud literally called his graph a "bell shape," though he had probably not encountered that term in a school situation before (three other students also described their graphs as bell shapes). This is probably a case of reinvention. Chris's graph was probably inspired by line graphs that the students made during mathematics lessons. She introduced the vertical axis with frequency, though such graphs had not been used before in the statistics course. Sandra probably started with the dots and then drew the continuous shape.

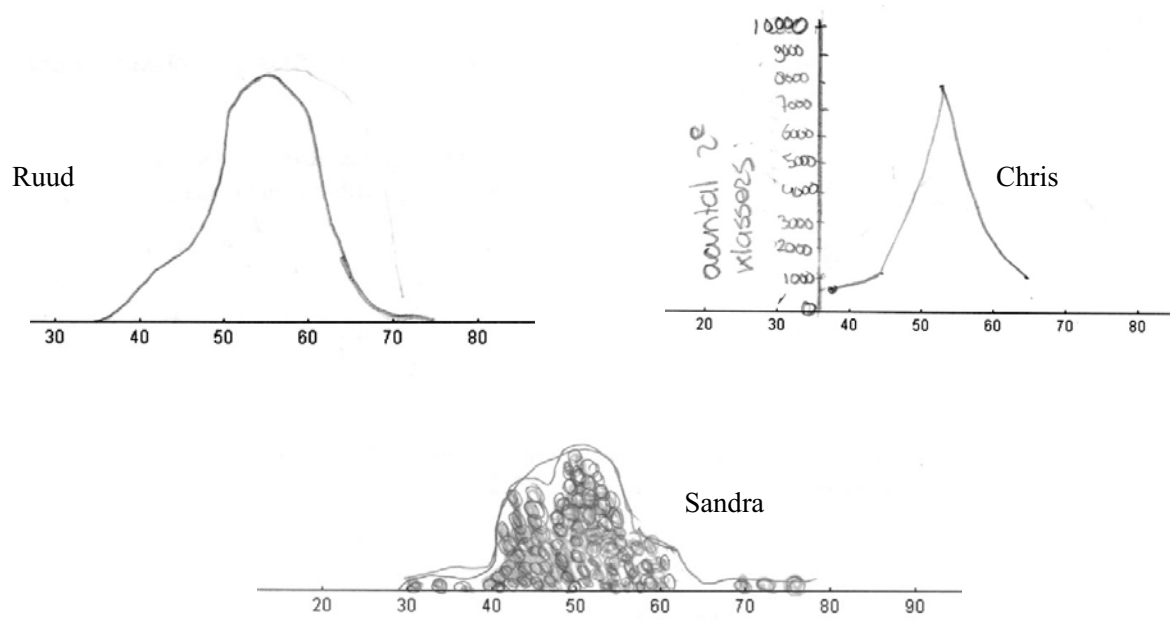

Figure 4. Predicted graphs for all students in the city by Ruud, Chris, and Sandra

In this third cycle of growing a sample, 23 students drew a bump shape. The words they used for the shapes were pyramid (three students), semicircle (one), and bell shape (four). Although many students draw continuous shapes, I did not exactly know what these shapes meant for them. Therefore, in the next section, I analyze students' reasoning with such shapes in the sixth lesson, which built on the fourth lesson. Furthermore, almost all student graphs looked roughly symmetrical, which is not surprising when the history of distribution is taken into account (Steinbring, 1980). In real life, however, the phenomenon of weight shows distributions that are skewed to the right. The skewness of weight data is caused by a "left wall effect" (two students had in fact drawn a left wall in the fourth lesson). By a left wall I mean that the lower limit (say about 30 kg) is relatively close to the average (53 kg) and the upper limit is relatively far away from the average (for example, sumo wrestlers can weigh 350 kg). The lower limit of 35 kg serves as a left wall, because adults can hardly live if they are lighter than 30 kg. This left wall in combination with no clear right wall causes the distribution to be skewed to the right. So far we had focused on spread and center as the core aspects of distribution, but skewness is another important characteristic of a distribution. Once there are different shapes to talk about, for example symmetrical or skewed, students can characterize shapes with different predicates. According to the hypothetical learning trajectory, skewness therefore had to become a topic of discussion as well in the following lessons. The next section shows how this was accomplished in the sixth lesson.

3.2. REASONING ABOUT SHAPES

In collaboration with the teacher, the following activity was designed with the purpose to make shape and in particular skewness a topic of discussion. To focus the students' attention on shape and skewness, the five shapes depicted in Figure 5 were drawn on the blackboard. They included three shapes mentioned by the students (a semicircle, a pyramid, a bell shape) and two skewed shapes (one unimodal distribution skewed to the right, and one skewed to the left). Students had to explain which shapes could not match the general distribution of people's weights based on their knowledge. The teacher expected that it would be easier for students to engage in the discussion if they could argue which shapes were not correct, instead of defending the shape they had chosen.

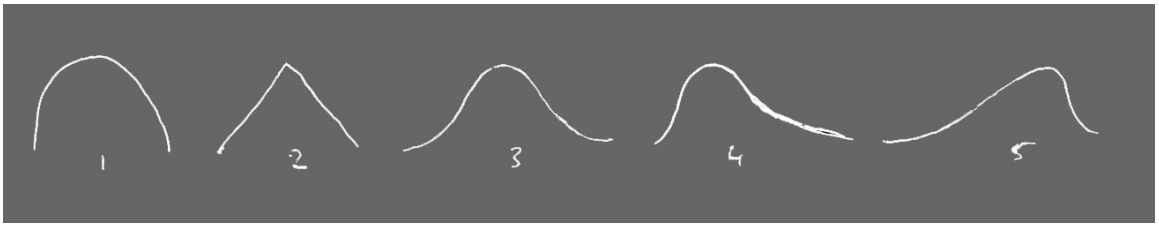

Figure 5. Five shapes as drawn on the blackboard: (1) semicircle, (2) pyramid, (3) normal distribution, (4) distribution skewed to the right, (5) distribution skewed to the left

The teacher chose students from the groups who thought that a particular shape on the blackboard could not be right. For all shapes except the normal shape, many students raised their hands. Apparently, most students expected a "normal" shape (number 3 in Figure 5).

1. First, Gigi explained why the semicircle (1) could not be right.

Gigi:	Well, I thought that it was a strange shape (...) For example, I thought that the average was about here [a little to the right of the middle] and I thought this one [top of the hill] was a little too high. It has to be lower. And I thought that here, that it was about 80, 90 [kg], and I don't think that so many people weigh that much or something [points at the height of the graph at the part of the graph with higher values].
Teacher:	(...) Does everybody agree with what Gigi says?
Tom:	Yes, but I also had something else. That there are no outliers. That it is straight and not that [he makes a gesture with two hands that looks like the tails of a normal distribution]. I would expect it to slope more if it goes more to the outside [makes the same gesture].

These students used statistical notions such as "outliers", although in an unconventional way, and height to explain shape issues, especially frequency. Furthermore, they used their knowledge of the context to reason about shape.

2. Because all of the students seemed to agree that the semicircle was not the right shape, the teacher wiped it off the blackboard and turned to the pyramid shape (2). This discussion involved "outliers" (the extreme values) and the mean in relation to shape.

Mourad:	Well, I didn't think this was the one, because, yeah, I don't think that a graph can be that rectangular.
Teacher:	The graph is not so rectangular? [inviting him to say more]
Mourad:	No, there are no outliers or stuff.
Alex:	It does have outliers; right at the end of both it does have outliers.
Student:	That is just the bottom [of the graph].
Alex:	At the end of the slanting line, there is an outlier, isn't it? (...)
Anna:	But the middle is the mean and everything else is outlier. [Other students say they do not agree, e.g. Fleur]
Fleur:	Who says that the middle is the mean?
Anna:	Yes, yes, roughly then.
Teacher:	Tom, you want to react.
Tom:	Look, if you have an outlier, then it has to go straight a bit [makes a horizontal movement with his hands]; otherwise it would not be

	an outlier (...) but that is not what I wanted to say. I wanted to react, that it [this graph] could not be the right one, because the peak is too sharp and then the mean would be too many of exactly the same.
Mike:	He just means that of one weight exactly all these kids have the same weight, so if the tip is at I-don't-know-how-many kilos, maybe 60 kilos, that all these kids are exactly 60 kilos.

This transcript shows that students started to react to each other. Before this lesson they mainly reacted to questions from the teacher, a type of interaction that is very common unfortunately (Van den Boer, 2003). In other words, the activity stimulated students to participate and their passive attitude started to change. Because the students agreed that the pyramid was not the right shape, the teacher wiped this shape off the blackboard also.

3. Next, Sofie was asked to explain why the bell shape (3) could not be the right shape. Before the discussion, almost all of the students thought this was the right shape (one girl admitted she did not know).

Sofie:	I had it that this was not the one, because there are also kids who are overweight. Therefore, I thought that it should go a bit like this [draws the right part a little more to the right, thus indicating a distribution skewed to the right, like Figure 5.4]. (...)
Rick:	That means that there are more kids much heavier, but there are also kids much less, so the other side should also go like that [this would imply a symmetrical graph].
Tom:	Guys, this is the right graph!

Because there was no agreement, the teacher did not wipe the graph off the board.

4. Next, Mike had to explain why he thought that the fourth, skewed graph could not be right.

> I thought that this was not it because... if the average is perhaps, if this it the highest point, then this [part on the left] would be a little longer; then it would have a curve like there [left half of the third graph]. I think that this cannot be right at all, and I also find it strange that there are so many high outliers. Then you would maybe come to 120 kilos or so. [Note that there were no numbers in the graphs.]

5. Last, Ellen spoke about the fifth graph, which was skewed to the left:

> Well, I think this one is also wrong because there are more heavy people than light people. And I think that eighth graders are more around 50 kilos. That's it.

Tom then objected, "it says 50 nowhere," and a lively discussion between the two evolved.

Thus, as intended, skewness became a topic of discussion, even in relation to center and "outliers". Next time we would certainly want to pay more attention to what students mean by such terms. Some students argued that the mean need not be the value in the middle. Still students seemed to make no clear distinctions between midrange, mean, and mode. Because the mode is not a measure that is often used in statistics, it was not the intention to address the mode unless students were

already reasoning with it. Since students at this point argued about the mean versus the value that occurred the most, the teacher and I decided to introduce a name for the mode, which these students had not learned before.

Researcher:	The value that occurs the most often has a name; it is called the mode [pointing at the value where the distribution has its peak]. (...) Who can explain in this graph [skewed to the right] whether the mean is higher or lower than the mode? (...)
Rick:	There are just more heavy people than light people, and therefore the mean is higher. (Note: Rick's remark makes sense if we interpret it to mean that heavy people are those to the right of the mode and light people those left of the mode).

In this way, there were opportunities to introduce statistical terms and relate them to each other, because students were already talking about the corresponding concepts or informal precursors to them. Traditionally, the mode is just introduced as the value that occurs the most, but here it was introduced as a characteristic of a distribution, albeit informally. The median was introduced later, in the ninth lesson, as the value that yields two equal groups (as can be done in minitool 2).

4. DISCUSSION

The main question of the overall research was: how can we promote coherent reasoning about variability, sampling, data, and distribution in a way that is meaningful for students with little statistical background? By carrying out several teaching experiments, some instructional activities turned out to be more effective than others. The activities analyzed in this paper, concerning growing a sample and reasoning about shapes, appeared particularly engaging and useful. The purpose of this paper was therefore to analyze students' learning process as exemplified in these two instructional activities and learn more about their development of concepts underlying statistical data analysis. I used Peirce's semiotics as an instrument of analysis in order to detect what the crucial elements of those activities were and what kind of learning these activities supported. The next section (4.1) discusses those key elements for each activity and speculates about what can be learned from the analysis presented here. The last two sections address the limitations (4.2) and implications (4.3) of the study.

4.1. THE INSTRUCTIONAL ACTIVITIES

Growing a sample (fourth lesson). The activity of growing a sample involved short cycles of constructing diagrams of new hypothetical situations, and comparing these with other diagrams of a real sample of the same size. The activity has a broader empirical basis than just the teaching experiment reported in this paper, because it emerged from a previous teaching experiment (Bakker & Gravemeijer, 2004) as a way to address shape as a pattern in variability and also resembles the growing samples activity described by Konold and Pollatsek (2002) in a fifth-grade classroom.

The activity was also based on design heuristics that were defined during previous teaching experiments (Bakker, 2004). One of those heuristics is to sometimes stay away from data so as to avoid students adopting a case-oriented view. Also, by asking students to compare their own diagrams with those representing real data, we invite them to "compare forests instead of trees"—compare data sets instead of individual data points. Moreover, by letting students predict a situation, the need is created to use conceptual tools for predicting that situation. By the cyclic approach taken such design heuristics for statistics education were validated by the research.

In the design of the activity several other issues also played a role. First, the teacher and I have wondered if the context of weight was suitable for this age group (13 years old). Many teachers and

textbooks avoid this context because it is so sensitive, but we found it striking how well students knew this context and how their predictions resembled the actual samples in many respects. The delicacy of this subject might explain part of their engagement during class discussions.

Another important pedagogical issue is the length of class discussions. In earlier lessons the teacher and I had noticed that these students found it hard to concentrate during class discussions for longer than about ten minutes. A cycle of producing a diagram for a sample of a specific size, and comparing it with a real sample requires only short periods of concentration. Providing real data in between their inventions demanded short periods of reflection and feedback. We also promoted more individual work than in a previous experiment so as to give all students the opportunity to predict and reflect themselves as opposed to listen to other students during one long class discussion.

As a way to generalize the results, I analyzed students' reasoning as an instance of diagrammatic reasoning, which typically involves constructing diagrams, experimenting with them, and reflecting on the results of the previous two steps. More generally, Bakker and Hoffmann (in press) argue that diagrammatic reasoning forms an opportunity for concept development. In this growing samples activity, the quick alternation between prediction and reflection during diagrammatic reasoning appears to create ample opportunities for hypostatic abstraction, for instance of the notion of spread.

In the first cycle involving predicting a small data set, students noted that the data were more spread out, but in subsequent cycles, students wrote or said that the spread was large. From the terms used in this fourth lesson, I conclude that many statistical concepts such as center (average, majority), spread (range and range of subsets of data), and shape had become topics of discussion (hypostatic abstractions) during the growing samples activity. Some of these words were used in a rather unconventional way, which implies that students needed more guidance at this point. Shape became a topic of discussion as students predicted that the shape of the graph would be a semicircle, a pyramid, or a bell shape, and this was exactly what the hypothetical learning trajectory (HLT) aimed at. Given the students' minimal background in statistics and the fact that this was only the fourth lesson of the sequence, the results were quite promising. Note, however, that such activities cannot simply be repeated in other contexts; they always need to be adjusted to local circumstances if they are to be applied in other situations.

Reasoning about shapes (sixth lesson). The aim of reasoning about shapes was that students would learn to reason about skewed shapes, and they did so in terms of the context (e.g. using "heavy" and "light"). The satisfactory outcome of this activity was that the students came to reason with statistical notions in a way they had not demonstrated before and were more engaged in the discussion than we had observed before, and this included students with low grades for mathematics.

To learn from the results, I speculate on the crucial features of this activity. First, the lack of formal rules and definitions probably makes it easier for low-achieving students to participate in the discussion. I furthermore conjecture that the lack of data, the game-like character and students' knowledge about the context were important factors, but also the fact that they had to argue *against* certain shapes. Such reasoning is safer than defending the shape they think is right. Note that the way in which mode, average, and other statistical notions were discussed contrasts drastically with what is common in statistics curricula because average values and spread were discussed in relation to shape, not just as computational operations on data values.

As mentioned in the beginning of the paper, I strove for a process of guided reinvention. This notion hints at the challenge of striking the balance between giving guidance to students on the one hand and giving them the freedom to reason using their own terminology on the other. This issue can be illustrated with a metaphor that Frege wrote to Hilbert in 1895 (Frege was one of the first modern logicians and philosophers of language, and Hilbert was a formalist mathematician). The topic was using and making symbols in mathematical discourse.

I would like to compare this with lignification [transformation into wood]. Where the tree lives and grows, it must be soft and sappy. If, however, the soft substance does not lignify, the tree cannot grow higher. If, on the contrary, all the green of the tree transforms into wood, the growing stops. (Frege, 1895/1976, p. 59; translation from German)

On the one hand, if statistical concepts are defined before students even have an intuitive idea of what these concepts are for (such as mean, median, and mode), then the tree transforms into wood and

students' conceptual development can be hindered (as discussed in the beginning of this paper). On the other hand, if teachers and instructional materials do not guide students well in a process of reinvention, the tree stays weak and cannot grow higher. It is evident that the notions of average, outliers, distribution, and sample in the present research needed to be developed into more precise notions, but at least students developed a language that was meaningful to them, an image that could be sharpened later on or, staying with the metaphor, a sappy part of the tree that can be lignified later. With reference to the lignification metaphor, the teacher and I had been reasonably successful in getting students to participate in reasoning about these shapes. However, they often used terms (in particular "outliers") in unconventional or vague ways, which is not surprising given the small number of lessons (ten in total) of the teaching experiment.

In terms of diagrammatic reasoning, this lesson was mainly devoted to reflection on shapes, but there were also examples of mental experimentation (what would the shape look like if...). Skewness was addressed within the weight context, but had not been predicated yet in terms of "left-skewed" or "right-skewed." Students mainly used two distribution aspects in their reasoning, average and the tails (what they called "outliers"). These notions are hypostatic abstractions that have become reasoning tools. From the analysis I concluded that students probably had the following understanding of distribution: there are many values around average (high rounded part in the sketch) and few low and high values, which is evidenced by the horizontal tails of the shape. This was indeed aimed at in the hypothetical learning trajectory.

4.2. LIMITATIONS

The purpose of analyses such as the one presented in this paper is that researchers and teachers can adjust such instructional activities to their own circumstances. A hypothetical learning trajectory always remains hypothetical, but others may learn from it, provided the conditions in which the design research has been conducted are clear to the audience. According to Freudenthal (1991, p. 161),

[design] research means: experiencing the cyclic process of development and research so consciously, and reporting on it so candidly that it justifies itself, and that this experience can be transmitted to others to become like their own experience.

It is therefore necessary to highlight the conditions and limitations of this study, which we do in this section.

Relevant information to judge the results in this paper is first that the teacher was experienced (11 years of teaching) and was preparing her dissertation in mathematics education. The other adults in the classroom were myself and assistants, who interviewed and observed students, and avoided teaching. Yet the mini-interviews probably had a learning effect, which means the interview questions we asked should be considered part of the HLT. The growing samples activity was successful in one seventh and one eighth grade class, but the activity of reasoning about shapes was only carried out in a single eighth grade class. Unlike with the growing samples activity we do not have more empirical support for the value of the reasoning about shapes activity than from this one class of 30 students. The school was not exceptional for a *havo-vwo*-school in the center of a large Dutch city, which implies that students probably belonged to the top 40% of Dutch students.

Researchers who would like to repeat such activities also need to take into account that we asked students about sample size from the first lesson onwards and that we tried to foster a classroom culture in which students were willing to discuss. This was not at all easy, because they were not used to whole-class discussion, but like Dutch students, used to self-reliant working on small tasks of a computational nature. We were therefore pleasantly surprised that a student, near the end of the teaching experiment, characterized statistics as "a little arithmetic and a lot of thinking."

As mentioned in the analysis of the growing samples activity, students' diagrams were strongly influenced by the two minitools they had used, but they also used line graphs taught in mathematics lessons. It is hard to decide whether the use of the minitools limited students' own diagrams to those provided by the software or whether it inspired them to make diagrams they would not have made without prior experience with the software. An argument for the former is that students did not make

any other type of graph than they had used with the minitools or in mathematics lessons. An argument for the latter is that the minitool representations were apparently meaningful to them despite the short exposure to them (by the third lesson they had only seen the second minitool, but had no hands-on experience with it). Other research designs such as comparative studies may be needed to decide such issues.

Unlike with the instructional materials developed for grade 7, the statistics unit for grade 8 has not been implemented in a school. Based on the experiences with two novice teachers who used the materials for grade 7, I expect that other teachers than the experienced teacher I worked with would need more time to reach similar results and it is possible that their students would not reach the same quality of reasoning in a first attempt to teach the unit without researchers and assistants interviewing in the classroom. The results presented in this paper therefore need to be interpreted as being possible to recreate given favorable conditions of sufficient time and support being available.

4.3. IMPLICATIONS

As a springboard to implications for teaching, assessment, and research, I raise the following question: why do almost all school textbooks follow the same routes and introduce mean, median, and mode as a trinity, and provide students with graphical tools such as histogram and box plot long before students have the conceptual understanding to use such tools sensibly? G. Cobb (1993, parag. 53) compared the situation with a night picture of a city: "if one could superimpose maps of the routes taken by all elementary books, the resulting picture would look much like a time-lapse night photograph of car taillights all moving along the same busy highway". Apart from the phenomenon of copying what others do, one important reason for this phenomenon could be that mean, median, mode, and graphs seem so easy to teach and, even more importantly, to assess. As argued in the beginning of this paper, however, this view easily leads to superficial understanding if students are not provided with ample opportunities to develop conceptual understanding of these statistical notions and graphs. The instructional activities presented in this paper are attempts to give students such opportunities.

Teaching. For several reasons, the approach taken in this paper is challenging for the teacher. The teacher plays an important role in steering the topic of discussion towards statistically important issues such as center and spread. This requires establishing a classroom culture in which students are willing to engage in discussions, which can be hard if they are used to working self-reliantly.

As argued before, the episodes analyzed in this paper can be framed as instances of diagrammatic reasoning, the key steps of which are making a diagram, experimenting with it, and reflecting on the results. During this diagrammatic reasoning, hypostatic abstractions such as majority, average, and shape can become objects and reasoning tools in the discourse. The analyses suggest the following recommendations that are not tied to the particular instructional activities considered here.

First, it is clear students need to diagrammatize—make their own diagrams that make sense to them, but also learn to make powerful conventional types of diagrams that are likely to become meaningful to them. The results show that the minitools software had a large influence on the diagrams students made themselves (see Section 4.2 for a discussion of this issue).

Second, students need to experiment with diagrams. Educational software such as the minitools can be useful in this stage of diagrammatic reasoning. The software should offer diagrams that students understand, but it should also offer opportunities for learning more advanced, culturally accepted diagrams. Apart from physical experimentation, mental experimentation is important, for instance when answering questions about hypothetical situations (Bakker & Gravemeijer, 2004).

Third, reflection should be stimulated. Throughout the research we noticed that the best reasoning occurred during teacher-directed class discussions that were not in the computer lab. One of the core issues of the reflection step is that students learn to describe ("predicate") and predict aggregate features of data sets, because that is an essential characteristic of statistical data analysis. Predicates should become topics of discussion so that they can be taken as entities in themselves. For example, talking about "most" data can lead to talking about the "majority"; describing how dots are "spread

out" can lead to saying that "the spread is large." These are examples of what Peirce called "hypostatic abstraction."

It is striking that these steps of diagrammatic reasoning, though they appear to be crucial to learning statistics, are so underexposed in most school textbooks. If we accept that diagrammatic reasoning is a basis for concept development (cf. Bakker & Hoffmann, in press), the above options are worth considering. A possible sequence that teachers could follow is (1) let students make their own diagrams but also offer types of diagrams that are likely to become meaningful, (2) enable students to experiment with diagrams both physically (e.g. using software) and mentally (e.g. by asking what-if questions), and (3) involve students in a reflection step in which they describe precisely what they see (clumps, majorities, shapes) and where they see it in a diagram.

Assessment. The learning that results from an approach of the form taken here may be harder to assess than whether students have learned to calculate average values or draw a histogram. In the approach taken here, the teacher has to accept that students' notions stay informal for a while, provided enough effort is taken to make informal notions more precise. In countries and states with high-stake accountability for the assessment of students' progress, this may be difficult to accomplish (cf. Makar & Confrey, 2004). We therefore need assessment items that assess what we find important and that might be used on large-scale tests.

Research. More research is needed into the question of how students can develop their own informal notions, such as center clumps, spread, and shapes, into conventional measures of center, variation, and other distribution aspects, and how teachers can support this development. The semiotic analysis suggests that one key issue is that the topic of a group discussion should be clear, and the teacher plays an important role in directing students' attention to that topic. Research is needed into the question of how teachers can be supported to help students in this regard.

ACKNOWLEDGMENTS

I thank Koeno Gravemeijer for his advice during the research, Corine van den Boer for teaching the eighth-grade class, and Carolien de Zwart, Sofie Goemans, and Yan-Wei Zhou for assisting in multiple ways. I also thank Nathalie Kuijpers for translating the transcripts and correcting the English, Jantien Smit, Anneleen Post, Katie Makar, and Phillip Kent for their editing help, and the editor Iddo Gal and anonymous reviewers for their helpful suggestions. The research was funded by the Netherlands Organization for Scientific Research under grant number 575-36-003B.

REFERENCES

Bakker, A. (2003). The early history of statistics and implications for education. *Journal of Statistics Education, 11*(1). [Online: www.amstat.org/publications/jse/v11n1/bakker.html]

Bakker, A. (2004). *Design research in statistics education: On symbolizing and computer tools.* Utrecht, The Netherlands: CD-Beta Press. [Online: http://www.library.uu.nl/digiarchief/dip/diss/2004-0513-153943/inhoud.htm]

Bakker, A., & Gravemeijer, K. P. E. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 147–167). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Bakker, A., & Hoffmann, M.H.G. (in press). Diagrammatic reasoning as the basis for developing concepts: A semiotic analysis of students' learning about statistical distribution. *Educational Studies in Mathematics*.

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

Cobb, G. (1993). Considering statistics education: A National Science Foundation Conference. *Journal of Statistics Education*, *1*(1). [Online: www.amstat.org/publications/jse/v1n1/cobb.html]

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

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

Cobb, P., Gravemeijer, K. P. E., Bowers, J., & Doorman, M. (1997). *Statistical Minitools* [applets and applications]. Nashville, TN & Utrecht: Vanderbilt University & Freudenthal Institute, Utrecht University. [Translated and revised in 2001 by Bakker, A.] [Online: www.wisweb.nl/en]

Cobb, P., McClain, K., & Gravemeijer, K. P. E. (2003). Learning about statistical covariation. *Cognition and Instruction 21*, 1–78.

Frege, F. L. G. (1976). *Wissenschaftlichter Briefwechsel [Scientific correspondence]* (Gabriel, G., Ed.) (First ed. Vol. 2). Hamburg, Germany: Meiner.

Freudenthal, H. (1991). *Revisiting mathematics education: China lectures.* Dordrecht, the Netherlands: Kluwer Academic Publishers.

Garfield, J., & Ben-Zvi, D. (2004). Research on statistical literacy, reasoning, and thinking: Issues, challenges, and implications. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 397–409). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory; Strategies for qualitative research*. Chicago: Aldine Publishing Company.

Gravemeijer, K. P. E. (1994). Educational development and developmental research. *Journal for Research in Mathematics Education, 25*, 443–471.

Gravemeijer, K. P. E. (1999, April). An instructional sequence of analysing univariate data sets. Paper presented at the Annual Meeting of the American Education Research Association, Montréal, Canada.

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

Konold, C., & Higgins, T. (2003). Reasoning about data. In J. Kilpatrick, W. G. Martin & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 193–215). Reston, VA: National Council of Teachers of Mathematics.

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

Konold, C., Robinson, A., Khalil, K., Pollatsek, A., Well, A. D., Wing, R., & Mayr, S. (2002). Students' use of modal clumps to summarize data. In B. Phillips (Ed.), *Proceedings of the International Conference on Teaching Statistics*, *Cape Town, South Africa*. [CD-ROM] Voorburg, The Netherlands: International Statistics Institute.

Makar, K., & Confrey, J. (2004). Secondary teachers' reasoning about comparing two groups. In D. Ben-Zvi & J. Garfield (Eds.), *The challenges of developing statistical literacy, reasoning, and thinking* (pp. 353–373). Dordrecht, the Netherlands: Kluwer Academic Publishers.

McGatha, M., Cobb, P., & McClain, K. (2002). An analysis of students' initial statistical understanding: Developing a conjectured learning trajectory. *Journal of Mathematical Behavior, 21*, 339–355.

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

Peirce, C. S. (1976). *The new elements of mathematics* (Eisele, C., Ed.) (Vol. I-IV). The Hague-Paris/Atlantic Highlands, N.J.: Mouton/Humanities Press.

Petrosino, A. J., Lehrer, R., & Schauble, L. (2003). Structuring error and experimental variation as distribution in the fourth grade. *Mathematical Thinking and Learning, 5*(2&3), 131–156.

Russell, S. J., & Corwin, R. B. (1989). Statistics: The shape of the data. *Used numbers: Real data in the classroom. Grades 4-6*. Washington, DC: National Science Foundation.

Schwartz, D. L., Goldman, S. R., Vye, N. J., Barron, B. J., & Cognition and Technology Group at Vanderbilt (1998). Aligning everyday and mathematical reasoning: The case of sampling assumptions. In S. P. Lajoie (Ed.), *Reflections on statistics: Learning, teaching, and assessment in grades K-12* (First ed., pp. 233–273). Mahwah, NJ: Lawrence Erlbaum Associates.

Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education, 26*, 114–145.

Steinbring, H. (1980). *Zur Entwicklung des Wahrscheinlichkeitsbegriffs - Das Anwendungsproblem in der Wahrscheinlichkeitstheorie aus didaktischer sicht.* [*On the development of the probability concept - The applicability problem in probability theory from a didactical perspective*]. Bielefeld: Institut für Didaktik der Mathematik der Universität Bielefeld.

Van den Boer, C. (2003). *Als je begrijpt wat ik bedoel. Een zoektocht naar verklaringen voor achterblijvende prestaties van allochtone leerlingen in het wiskundeonderwijs* [If you know what I mean. A search for an explanation of lagging results of mathematics education among ethnic minority students]. [CDROM] Utrecht, The Netherlands: Beta Press.

Van Oers, B. (2000). The appropriation of mathematics symbols: A psychosemiotic approach to mathematics learning. In P. Cobb, E. Yackel & K. McClain (Eds.), *Symbolizing and communicating in mathematics classrooms; Perspectivies on discourse, tools, and instructional design* (pp. 133–176). Mahwah, NJ: Lawrence Erlbaum Associates.

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

ARTHUR BAKKER
Since September 1, 2004:
Institute of Education
University of London
23-29 Emerald Street
London WC1N 3QS
United Kingdom

STUDENT DESCRIPTION OF VARIATION WHILE WORKING WITH WEATHER DATA

CHRIS READING
University of New England
creading@metz.une.edu.au

SUMMARY

Variation is a key concept in the study of statistics and its understanding is a crucial aspect of most statistically related tasks. This study aimed to extend and apply a hierarchy for describing students' understanding of variation that was developed in a sampling context to the context of a natural event in which variation occurs. Students aged 13 to 17 engaged in an inference task that necessitated the description of both rainfall and temperature data. The SOLO Taxonomy was used as a framework for analyzing student responses. Two cycles of Unistructural-Multistructural-Relational levels, one for qualitative descriptions and the other for quantitative descriptions, were identified in responses. Implications of the extended hierarchy for describing understanding of variation for research, teaching and assessment are outlined.

Keywords: Describing variation; SOLO Taxonomy; Inference task; Secondary students

1. INTRODUCTION

The analysis of variation, that is, the irregularities in data, is critical to the study of statistics (Wild & Pfannkuch, 1999, p. 235). Despite this critical nature of variation, not much is known about how students perceive variation. A prior review of the literature has shown that, despite the importance of variation, most research examines the understanding of central tendency and that research on understanding of variation is limited (Shaughnessy, Watson, Moritz, & Reading, 1999). In fact, the work by Shaughnessy et. al (1999) is one of the first attempts to unpack, in a systematic way, what is happening in students' understanding of variation. Given that variation is critical to the study of statistics, more research needs to be undertaken to better understand how students view and describe variation. This study was undertaken to develop a hierarchy to assess students' understanding of variation. The results of this study are expected to assist researchers and teachers by providing a tool for describing the level of statistical sophistication in the description of variation.

1.1. STUDENTS' PROPENSITY TO DISCUSS VARIATION

When dealing with data, consideration needs to be given to both measures of central tendency and measures of variation. So, which of these are students more likely to use if not prompted when working with data? Research has shown that when engaged in reducing data, although some students base their responses on measures of variation, many more students use measures of central tendency (Reading & Pegg, 1996, p.190). This investigation involved Australian secondary school students for whom most data reduction learning experiences deal with finding 'mean, mode and median', hence it is not surprising that so few bother with measures of variation. On the other hand, in Australian schools students are presented with few learning experiences that involve making inferences from data and generally are not given specific instruction as to how to engage in such activities. Thus, responses to tasks that involve making inferences are less likely to reflect approaches imposed by teachers. In fact, analysis of secondary school student responses to open-ended questions involving

making inferences from data showed that more students based an inference discussion on measures of variation than measures of central tendency (Reading, 1998, p. 1430).

The conflicting student inclinations in these two different experiences, engaged in reducing data and engaged in making inferences about data, suggest that students may have a propensity to consider measures of variation when dealing with data but unless this is given a chance to develop, such a propensity may eventually be overcome by the 'push from teachers' to discuss measures of central tendency. Even more importantly though, if teachers concentrate their efforts on working with measures of central tendency, then students will be denied the opportunity to experience situations where they can begin to understand variation and to develop any propensity they may have to reason about variation. Already researchers are recognizing the need to develop learning situations where students can be encouraged to develop the notion of variability. One such approach is Bakker's (2003) 'growing samples' activity that allows students to investigate the shape of distributions as a basis for developing a better understanding of variability. However, the present research was not designed to determine whether early attempts at inference by students are more likely to be based on measures of central tendency or variation but, to consider aspects of reasoning about variation that do become apparent when students make inferences.

1.2. CONSIDERATION OF VARIATION

The study of measures of variation in schools, such as the standard deviation, has developed notoriety with teachers as being particularly cumbersome, resulting in many teachers having difficulty developing the concept with students or avoiding it altogether. This is unfortunate given that many students show, at least in some contexts, a natural propensity to base discussion of data on measures of variation rather than central tendency (Reading, 1998). In order to be better prepared to equip students with an understanding of variation, teachers need to understand how students reason about variation and also to have a means for assessing how students reason about variation.

Concern over lack of attention to variation has prompted researchers to investigate in more detail students' understanding of variation. Some studies were undertaken following dissatisfaction with the responses of Grade 4 students in the USA to a National Assessment of Educational Progress (NAEP) test item (Shaughnessy et al., 1999; Reading & Shaughnessy, 2000). In this extended response item, students had to predict the number of red gum balls in a sample of ten obtained from a gum ball machine, and then explain their reason(s) for choosing that number. This task allowed students to demonstrate their understanding of *centrality* (expected outcome according to formal probability calculation) but not *variability* (in outcomes across repeated trials). Shaughnessy et al. (1999) redesigned this task and analyzed pencil and paper responses to gain useful information about students' conceptions of variation in a similar sampling situation based around a candy bowl rather than a gum ball machine. These modified investigations have been extended in various contexts (Torok & Watson, 2000). In particular, Shaughnessy and Ciancetta (2002) allowed students to experience the variability in results, with ten trials of a spinner task, before predicting the outcomes.

When outlining the foundations of 'thinking statistically' Wild and Pfannkuch (1999, p. 226) identify 'consideration of variation' as one of the fundamental types of thinking. They list four components of consideration of variation: noticing and acknowledging variation; measuring and modelling variation for the purpose of prediction, explanation or control; explaining and dealing with variation; and developing investigative strategies in relation to variation. Reading and Shaughnessy (2004) have also suggested two additional components: describing variation; and representing variation. To best investigate students' reasoning about variation it is necessary to delve into as many as possible of these components, and examine how students describe the variation they observe and endeavour to interpret and/or use for inference.

2. DEVELOPMENTAL HIERARCHIES

The increasing popularity of research into cognitive frameworks to assess students' understanding of phenomena when responding to learning or assessment activities has provided the impetus for the

creation of developmental hierarchies in stochastics. Following is an introduction to a particular model for explaining developmental growth and then a summary of aspects of existing developmental hierarchies that particularly address variation.

2.1. THE SOLO TAXONOMY

Research into Developmental-Based Assessment (DBA), the assessment of students based on the quality of their understanding and learning (Pegg, 2003), has contributed to the increased acceptance of developmental frameworks. This approach to assessment, which focuses on the mental structure of understanding, differs from outcomes-based assessment which focuses on what students are expected to know. This paper focuses on the Structure of the Observed Learning Outcome (SOLO) Model, an approach to assessment which rests on an empirically established cognitive developmental model (Pegg, 2003).

The neo-piagetian SOLO Taxonomy (Biggs & Collis, 1991) consists of five modes of functioning, with levels of achievement identifiable within each of these modes. The two modes relevant to the present research are the ikonic mode (making use of imaging and imagination) and the concrete symbolic mode (operating with second order symbol systems such as written language). Although these modes are similar to Piagetian stages, an important difference is that with the SOLO Taxonomy earlier modes are not seen as replaced by subsequent modes and in fact are often being used to support growth in the later modes.

A series of levels of increasing cognitive development has been identified within each of these modes. The three levels relevant to this study are: *unistructural* responses - with focus on one element, *multistructural* responses - with focus on several unrelated elements, and *relational* responses - with focus on several elements in which inter-relationships are identified. These three levels form a cycle of cognitive growth, from unistructural, through multistructural, to relational responses, that occurs within a mode. For example, when describing a geometric figure, students may focus on an element such as a 'property of the figure'. Unistructural responses would describe one property of the figure, perhaps focusing on the lengths of the sides. Multistructural responses would address more than one property, perhaps the lengths of the sides and sizes of the angles. Relational responses would identify links and deal with a relationship between the properties, perhaps stating that adjacent angles being right angles would imply pairs of parallel sides. The relational level response in one cycle is similar to, but not as concise as, the unistructural response in the next cycle. Early applications of SOLO only described one cycle of levels within each mode, but more recently researchers have identified more than one cycle of levels within a mode (Pegg, 2003, pp. 244-245). This taxonomy is particularly useful because of the depth of analysis that can be achieved when interpreting students' responses.

2.2. DEVELOPMENTAL HIERARCHIES FOCUSING ON VARIATION

Neo-Piagetians have provided a foundation of cognitive frameworks on which to base developmental hierarchies in probability (e.g., Jones, Langrall, Thornton & Mogill, 1997) and in statistics (e.g., Mooney, 2002). SOLO has already been employed to explain statistical thinking frameworks (e.g., Jones et al., 2000; and refined by Mooney, 2002) and is more recently being used as the basis for development hierarchies related to variation (e.g., Watson, Kelly, Callingham & Shaughnessy, 2003). Mooney (2002) developed four SOLO based 'levels' in each of four processes. Variation is only mentioned in one of the four processes, 'organizing and reducing data', and the relevant descriptors for that process (Mooney, 2002, pp. 36-37) are reproduced in Table 1. A series of studies reported by Jones, Mooney, Langrall and Thornton (2002) was used to validate the SOLO-based levels.

Table 1. Partially reproduced Statistical Thinking Framework (Mooney, 2002)

	Organising and Reducing Data
Levels	**Focus of Responses**
1 - Idiosyncratic	Is not able to describe the spread of the data in terms representative of the spread.
2 - Transitional	Describes the spread of the data using invented measures that are partially valid.
3 - Quantitative	Describes the spread of the data using a measure from a flawed procedure or a valid and correct invented measure.
4 - Analytical	Describes the spread of data using a valid and correct measure.

Watson et al. (2003) used the Torok and Watson (2000) hierarchy levels, in conjunction with SOLO, as a starting point for the analysis of responses to a bank of assessment items, culminating in the description of four levels for the understanding of statistical variation (Watson et al., 2003, p. 11) described in Table 2. Although these four levels were developed to measure understanding of variation they do not explain how students actually describe the variation.

Table 2. Developing Concepts of Variation (Watson et al., 2003)

Levels	**Focus of Responses**
1 - Prerequisites for variation	Working out the environment, table/simple graph reading, intuitive reasoning for chance.
2 - Partial recognition of variation	Putting ideas in context, tendency to focus on single aspects and neglect others.
3 - Application of variation	Consolidating and using ideas in context, inconsistent in picking most salient features.
4 - Critical aspects of variation	Employing complex justification or critical reasoning.

Reading and Shaughnessy (2000) interviewed 12 students regarding the sampling task used earlier by Shaughnessy et al. (1999) and identified non-sophisticated discussion of variation in this context. These interviews were further analyzed and two hierarchies for understanding of variation, one for description and the other for causation, were developed (Reading & Shaughnessy, 2004) based on students' perceptions in the sampling situation. Only the Description Hierarchy is relevant to the present study and the four levels of this hierarchy are summarized in Table 3.

Table 3. Description Hierarchy (Reading & Shaughnessy, 2004)

Levels	**Focus of Responses**
D1 - Concern with Either Middle Values or Extreme Values	Describe variation in terms of what is happening with either extreme values or middle values. *Extreme Values* are used to indicate data items that are at the uppermost or lowest end of the data, while *Middle Values* indicate those data items that are between the extremes.
D2 - Concern with Both Middle Values and Extreme Values	Describe variation using both the extreme values and what is happening with the values between the extremes.
D3 - Discuss Deviations from an Anchor	Describe variation in terms of deviations from some value but either the anchor for such deviations is not central, or not specifically identified as central.
D4 - Discuss Deviations from a Central Anchor	Describe variation by considering both a centre and what is happening about that centre.

The Reading and Shaughnessy (2004) Description Hierarchy was based on student responses to a sampling task but did not use SOLO as a conceptual framework. The present study was designed to use, and modify or extend if necessary, this hierarchy to code responses given in a weather-related inference task and to consider SOLO as a suitable conceptual framework to explain the hierarchy. A weather-based task was chosen because weather is a phenomenon which involves variation that everyone experiences, hence it can provide students with a meaningful context for data description and inference.

The main research question was how students describe variation during an inference task. This necessitated investigation of three related research questions: Is the hierarchy developed for analyzing students' descriptions of variation in a sampling situation (Reading & Shaughnessy, 2004) also applicable for coding responses with data descriptions given when making inferences from weather-related data, in which there is natural variation? If this hierarchy is suitable, does SOLO offer a broad framework for explaining the hierarchy? If SOLO is a suitable framework, can cycles of levels be identified within the SOLO modes? The findings would contribute to the refinement of conceptual models developed in earlier research, and could assist researchers and teachers by providing a developmentally-based hierarchy.

3. METHOD

This section describes an exploratory study that involved posing to students in Grades 7, 9 and 11 a weather-related inference task with two separate segments. The following describes the task students faced, procedure, participants, and analytic approach and associated issues.

3.1. WEATHER ACTIVITY – STUDENT TASK

The Weather Activity presented students with a Scenario, see Figure 1, based around choosing the most suitable month for a proposed Youth Festival to be held in the students' own town. It was stressed to students that they did not have to worry about any other aspects of the celebration, only the weather. The activity was implemented over a period of time and incorporated both data description and inference components.

WEATHER ACTIVITY SCENARIO

XXXXXX is to introduce a new celebration into the calendar. *Youth Alive* will celebrate the youth of the city and be held at an outdoor venue. Although not all details have been decided concerning the activities to be held on the day, a decision needs to be made as to a date for the celebration so that it can be slotted into the calendar. Organisers have expressed concern as to the effect that XXXXXX's often unpredictable weather could have on such a celebration. You have been commissioned to submit a report to the committee describing XXXXXX's weather and to suggest a suitable month for the celebration. Other factors will be taken into consideration to decide exactly which day in the month *Youth Alive* will be held.

Figure 1. Weather activity scenario

The activity was designed to have two separate segments, the first segment based around rainfall data and the second segment based around temperature data. Before each segment began, each student was randomly allocated one particular month of data to consider. The data used by the students in the task consisted of rainfall figures (daily millimetres) for 36 months in the first segment, and temperature figures (daily minimum and maximum temperatures in degrees celsius) for 36 months in the second segment. The use of 36 different months ensured that each student in the class had different data. The weather was chosen for the three years 1998 through to 2000 because the activity was undertaken in 2001. Examples of the data as presented to the students appear in the Appendix. It should be noted that the monthly data provided to each student exhibited different patterns and different variability.

Within each segment students had an individual task followed by a group task that were both open ended. First, each student was asked to individually describe the weather in his or her month in a written response. The students were told that these descriptions were to be used in the next step to compare with the descriptions provided for other months by a small group (about four) of classmates and decide on the most suitable month of the year for the festival from amongst those months within the group. Later, the students worked together in these groups, comparing their data descriptions, and developing a written group response that both described what they chose as the most suitable month for the festival, out of those they compared, and explained their reasoning.

The use of open-ended tasks meant that students had the freedom to adopt criteria or attend to issues they considered necessary. No specific instruction was given to discuss variation, despite the fact that description of variation was the focus of the research. This approach was taken so that students were free to discuss variation how and when they saw the need. This methodological approach has been utilized by other researchers. Watson et al. (2003) designed items to allow students to be free to demonstrate their understanding of variation and Ben-Zvi's (2003) open learning activity gave no specific direction to discuss variability despite the fact that analysis of the data was to include how students reasoned about variation.

3.2. PROCEDURE

The weather activity task was implemented in classes, during normal teaching time by a research assistant. Allocation of data sets to students was random but allocation of students to groups was not. The random allocation of data meant that students did not necessarily receive the same month for the temperature segment as they had used for the rainfall segment. The allocation of students into groups for the group activity was made by the teacher, based on knowledge about good working relationships from previous group work. If students were unsure about the task and needed a prompt, then it was suggested that they should look for any pattern in the data or any key features that may be useful.

The weather activity was planned to spread over a number of weeks to suit the school schedule. The individual and group work for each segment was planned to occur during one standard lesson time-slot for the class. The temperature segment lesson, however, did not immediately follow the rainfall segment lesson. Between the two segments of the activity there was a teaching episode, presenting a statistics-related section of the curriculum. These teaching episodes were requested by the teacher to align the weather activity with the students' learning experience. The episodes were implemented by the class teacher and involved demonstrating to the class statistical tools that might be useful when describing the data. Grade 7 students were introduced to stem-and-leaf plots and the summary statistics: maximum, minimum, average, and range. Grade 9 students were introduced to box-and-whisker plots, from a development base of stem-and-leaf plots with which they were already familiar. In Grade 11 an entire unit of work on statistics, including stem-and-leaf plots and box-and-whisker plots, was implemented between the two segments of the activity.

The weather activity provided both individual and group written responses for analysis. However, only the analysis of the individual responses will be reported here. For discussion of the group responses see Reading and Lawrie (2004). A wrap-up activity planned as the last segment of the weather activity, having individual students make a final decision about the most appropriate month with all data discussions available to them from all groups, was not completed because there was insufficient time due to the intervention of other events in the school. However, this did not detract from the usefulness of the responses provided in each of the rainfall segment and temperature segment as the students did not know at that point in time that the wrap-up activity would not be completed.

It is acknowledged that students could have developed their understanding regarding variation during the group discussion part of the rainfall segment and the teaching episode, in ways which could affect the quality of their later responses to the first (individual) part of the temperature segment. However, as the research was attempting to refine a hierarchy for coding responses and not to assess a student's performance at any particular instant or to compare performance before and after instruction, the possible improvements in quality of response in fact would provide a richer array of

data for refining the hierarchy. This sequence of activities enabled the researchers to capture the reasoning of a heterogeneous set of students at different points during their exposure to data with natural variation.

3.3. PARTICIPANTS

This research targeted students in Grades 7, 9 and 11 (aged 13 to 17) in a secondary school in a rural city in northern New South Wales, Australia. Students from one class in each of the three grades were included in the study. Classes were selected so as to include students with average mathematical ability. Only two teachers from the school were involved, as one teacher had charge of two of the classes. The actual number of students who completed each of the individual steps of the research activity was not consistent, as attendance in each class varied over the particular days when activities were presented. The breakdown of students participating in the weather activity is presented for the rainfall segment in Table 4 and for the temperature segment in Table 5.

Table 4. Participants in the rainfall segment of the weather activity

	Grade 7	Grade 9	Grade 11	Total
Male	15	17	7	39
Female	6	11	9	26
Total	21	28	16	65

Table 5. Participants in the temperature segment of the weather activity

	Grade 7	Grade 9	Grade 11	Total
Male	16	15	9	40
Female	5	9	10	24
Total	21	24	19	64

3.4. ANALYTIC APPROACH AND ASSOCIATED ISSUES

The purpose of the analysis of the individual responses was to determine the applicability of the Reading and Shaughnessy (2004) hierarchy, developed for a sampling situation, to coding responses to an inference task and refine the hierarchy if needed. Coding of the written discussions in the responses was undertaken in three stages. First, the responses were coded independently by the researcher and the research assistant, based on the Reading and Shaughnessy (2002) hierarchy in Table 3. Second, level descriptions were revised based on any newly identified descriptions of features of variation and the hierarchy was expanded by developing new levels based on responses not suitably accommodated. Such a revision process for coding hierarchies has been utilized by a number of researchers (e.g. Mooney, 2002; Langrall & Mooney, 2002; Watson et al., 2003). Finally, all responses were recoded independently by both the researcher and research assistant based on the new hierarchy. The recoding produced an 85% agreement and then discussion was used to resolve disputed codings. Such discussions also helped to refine the clarification of each level in the hierarchy. Before describing the hierarchy in detail some relevant aspects of student performance on the task are outlined.

Students were given access to graph paper but were not specifically required to produce a graph. While a number of students in each grade (especially in Grade 7, as shown in Table 6) chose to draw a graph as part of their response, only three responses (one Grade 9 and two Grade 11) actually referred to the information in the graph as part of their written explanation. All three responses were given during the temperature segment, after the teaching episode that involved graphing.

Table 6. Student creation of graphs

	Grade 7	Grade 9	Grade 11
Rainfall	67% (14/21)	21% (6/28)	26% (5/16)
Temperature	78% (13/21)	8% (2/24)	0% (0/19)

Almost all students included aspects of both a description and a prediction in their individual responses, rather than just a description as requested in the individual task. This compulsion to predict may reflect a need to give a purpose, predicting, to justify having to give a description. The discussion includes aspects of the responses irrespective of whether the student gave a prediction as well as description, or just gave a description as requested. Noticeable in the responses was whether the explanations were based on the given data and how much of the data were used.

Although many students mentioned factors external to the data when justifying decisions, most of them also referred to the data. Some external factors mentioned were of a personal and less relevant nature, such as the occurrence of a birthday, while others were of a less personal and more relevant nature, such as prior knowledge of local weather or similar events that have been held in the past. Encouragingly 72% of responses from Grade 7 students, 96% from Grade 9 and 97% from Grade 11 made at least some reference to the supplied data indicating that most students in the higher grades appreciated the need to use the data provided as a basis for the written explanations.

When describing the weather for the month some students used just some of the data by choosing to focus on a particular part of the month while others incorporated all of the month. Those focusing on part of the month generally chose specific day(s) or a consecutive sequence (block) of days. Many students focused on such a block when discussing the rainfall. This focus appears to have been influenced by the rain/no rain (dichotomous) nature often attributed to the rainfall variable. Focus on the whole month was more typical for the responses dealing with the temperature data and justifications for decisions often dealt with quoting one or more simple statistic(s), such as the maximum, minimum or average, for both sets of data as if they were one.

The references to features of variation in the data varied considerably in length and quality and were both qualitative and quantitative in nature. The following details the results of the analysis of the responses.

4. RESULTS

This study aimed to develop a way of assessing students' descriptions of variation, by extending the Reading and Shaughnessy (2004) Description Hierarchy presented earlier in Table 3. In the first step of the analysis, attempts were made to code the responses using the four levels D1, D2, D3 and D4 of this original hierarchy. Most responses were found to fall within the D1 and D2 levels, describing variation using *Extreme Values* or *Middle Values* or both. It soon became apparent, however, that while the written discussions were clearly falling into these levels, some were expressing the features in words only while others were expressing the features numerically. Those using words only, with no numeric descriptions of features of the variation, were labeled qualitative responses while those that did include numeric features were labeled quantitative. Thus the revision and expansion of the Reading and Shaughnessy (2004) hierarchy focused on developing two distinct groupings of responses based mainly on the previous D1 and D2 levels, one grouping based on qualitative descriptions of variation and the other based on quantitative descriptions. These two types of responses are analyzed separately in sections 4.1 and 4.2 below. Further discussion of the comparison of the Reading and Shaughnessy (2004) hierarchy and the proposed groupings of coded responses can be found in the Discussion section later on.

For easy reference in the discussion, an identification tag has been assigned to reproduced responses, based on consecutive appearance. The tags begin with R, followed by a grade number (7, 9 or 11) and then a specific student number. For example, R1102 is the second response from a Grade 11 student to appear in the discussion. Any response, or part thereof, that is reproduced directly is

shown in italics. No grammatical corrections have been made to the responses. When interpreting any reproduced responses it must be remembered that each student was dealing with data for a different month.

4.1. QUALITATIVE RESPONSES

The responses in this first grouping use word-based expressions, rather than numerical expressions, to describe the variation in the data. Some of these responses describe the variation by the use of general terms or phrases to describe the nature of the changes identified while others are more specific in their qualitative descriptions. Next is a discussion of the structure of responses at each of the three levels, unistructural, multistructural, and relational.

Unistructura1 Responses

The unistructural responses give one qualitative description to summarize an impression of the variation and can be grouped into two types, magnitude-related and arrangement-related. The magnitude-related terms, typical examples given in Table 7, are used in an absolute sense to give an indication of how the magnitude of the numbers is changing. Some terms suggest little change, while others suggest more change. The arrangement-related terms, typical examples given in Table 8, are used in a relative sense to give an impression of the position of the data elements relative to other data elements. Some terms suggest an inability to decide on any basis for the arrangement while others suggest a regular, describable arrangement. The use of the term *distributed*, see Table 8, is noteworthy. The students involved would most likely not have met the term 'distribution' in a formal statistical sense and these references may be in a more general sense of things being arranged.

Table 7. Magnitude-related phrases used in unistructural responses

Suggesting Little Change	Suggesting More Change
slightly on and off	*least predictable*
reasonably steady	*a bit unpredictable*
most consistent	*seem to be more mixed around*
pretty much consistent	*a bit erratic*
no sudden variations	*very unpredictable*
pretty regular	

Table 8. Arrangement-related phrases used in unistructural responses

Undecided on Arrangement	Decided on Arrangement
no pattern	*spread out*
no particular pattern	*scattered through*
no real pattern	*evenly spread*
	even balance
	evenly distributed
	distribution is limited

Multistructural Responses

The multistructural responses make use of more than one qualitative statement when describing the variation and fall into two categories, *Limiting* and *Sequential*. The first category, *Limiting*, comprises responses that deal with the data by setting a general limit on the values, often indicated by too much or too little. Such responses were more common for description of the temperature and typically summarize the data using the term 'too', such as *too cold* or *too hot*, for example R1101 and R901.

(R1101) *September wouldn't be good because its too cold in the morning…*

(R901) *It could be a good month to have it because it doesn't get too hot.*

Such responses were not as common for rainfall, but R701 and R702 are examples, with R702 qualifying the *just about nothing* description of the rain by giving the total.

(R701) *I think they should have it in february because theres not much rainfalls ...*

(R702) *I think my month is the best because it rains just about nothing at all. This months complete rainfall is 8.6 mm.*

The second category, *Sequential*, comprises responses that deal with the data item by item or by grouping like data items in a qualitative way. Such responses were more common for description of the rainfall and typically summarize the data into like blocks of days. These blocks tended to be wet days (rain) and dry days (no rain), reflecting the dichotomous nature imposed on the rainfall variable by students. Response R902 has generalized this blocking while R1102 is more specific about the blocks. Another form of description of this rain/no rain dichotomy was by pairing, such as in R903. The less common sequential temperature responses block off the days in the month based generally on a higher versus lower temperature dichotomy, as typified by R904.

(R902) *There are a lot of dry days then a couple of wet days then a lot of dry days again.*

(R1102) *In the first 10 days of the month would be good as there is no rainfall here and then it continues as 4 days with rain, 5 days clear, 2 days with rain, 7 days clear, 2 days with rain.*

(R903) *Usually the rain is in pairs. After a high column little or no rain is after it.*

(R904) *In the minimum temperatures there seems to be a pattern of a few higher temperatures and then a few lower temperatures and so on.*

Relational Responses

The relational responses give a qualitative description of the variation suggesting that both limiting and sequential aspects have been considered and linked to give an overall description. An example is R905, which gives a general limit but then goes on to discuss blocks in a sequential manner. Such linked responses, however, were uncommon for those who gave qualitative descriptions.

(R905*) January '98 seems to be a pretty average month in terms of rainfall. Not too much and not too little. The rain seems to fall pretty regularly, but the amounts are not much. I think January would be a good month to hold "Youth Alive". The main pattern seems to be a short spell of dry days (3-5 days) and then 1 or 2 wet days but as the rain is pretty light and not a large amount falls, I think this month would be pretty good.*

4.2. QUANTITATIVE RESPONSES

The responses in this second grouping use numerical values, often simple statistics, to describe the variation in the data. Next is a discussion of the structure of responses at each of the three levels, unistructural, multistructural, and relational.

Unistructual Responses

The unistructural responses discuss one quantitative feature when describing the variation and fall into two distinct categories, one based on a description of the *Extreme Values* of the data and the other based on *Interior Values*. The *Extreme Values* responses describe the extreme values of the data explicitly by referring to the minimum and/or maximum or implicitly by referring to the highest and/or lowest. Responses mentioning the maximum and/or minimum explicitly are easily identifiable, so the following examples particularly demonstrate some of the more implicit references. Few of these more quantitative responses gave only one extreme for the data, thus it was rare to find the minimum without the maximum and vice versa. *Extreme Values* responses were much more common with the temperature data, typified by R906, than rainfall, typified by R907.

(R906) *August 98 was a relatively cold month, the highest temp being only 17.9 degrees Celsius and the lowest being a freezing -6.9 degrees Celsius.*

(R907) *I can see that the highest rainfall was 45.2 ml and the lowest was 0.0ml.*

Many responses, such as R906, gave the highest and lowest data values for the temperature by quoting the highest figure for the Maximum Temperature and the lowest figure for the Minimum Temperature, as if the two separate variables were being treated as one temperature variable. It is possible, though, that students only considered the top of the maximums relevant to the task at hand and the bottom of the maximums not so relevant. Similarly, the bottom of the minimum temperatures could be considered more relevant than the top of the minimums. The response R703 gave, not just the *most* but also, the second highest rainfall for the month, before reverting to describing the total rainfall for blocks of days. Other ways of expressing the minimum or maximum implicitly included *doesn't go below* (see R704), *decreasing past* and *exceeding* (see R908).

(R703) *The 4th had the most rain with 31 mils second was the 5th with 29.8 mils so…*

(R704*) March would be a pretty sweet month to have this thingy in because it doesn't go below 5 degrees and usually about 30 degrees Celsius at peak temperature.*

(R908) *February doesn't seem to have a pattern except that it seems to have a fairly warm to hot climate with temperature either exceeding 30 degrees or decreasing past 4 degrees.*

A natural progression for those giving both the maximum and minimum was to describe the range. Some responses, such as R705, actually expressed the maximum and minimum in a *from... to...* form, thus supplying a maximum and minimum and implying a range. Other responses explicitly mentioned the range, either for just one of the variables, as in R1103, or for both, as in R1104.

(R705*) In between the 3rd and 12th would be a good time to have the thing with it been warm but not to hot. The max temps where 21 degrees Celsius to 29 degrees Celsius and the min temps where 5 degrees Celsius to 15 degrees Celsius.*

(R1103) *The maximum temp in March was 28.7 degrees Celsius the minimum was 5.6 degrees Celsius. The range in Max. temp was 12.3 degrees Celsius. I think this month would be good to hold the youth fest in because it stays fairly warm throughout the month.*

(R1104) *The maximum temperature average is 16.6 degrees Celsius which is cool but not too cold temperature. I don't think this max temp would be ideal for the youth fest. The max temps range from 14 - 19.8 degrees Celsius so the max highest is what I would be wanting. The min deg C ranges a lot from 10.5 - -0.5, this is cold weather and wouldn't suit a festival…*

The *Interior Values* responses describe the interior values of the data by referring to blocks of rainfall or temperature. Those responses mentioning the blocks were generally descriptions for rainfall. Sometimes the responses referred to blocks in general, as in R1105, while others were more specific about the number of days or the exact dates when they occurred, as in R1106. Some other ways of referring to the blocks included as patches (see R1107) and periods (see R706).

(R1105) *It appears that after a larger rainfall of 36mm it rains slightly on and off for the following week before another large rainfall. It also rains a few days before the heavy rainfall sort of like a build up and dies down at the end of the 2nd heavy rainfall.*

(R1106) *There seems to be rain nearly every 5 days for 1 - 3 days either side of the 5th day. The 18th seems to be the best day because it is in the middle of 15 and 20 and it is the middle day of a six day dry spell.*

(R1107*) In June 99 the temprature is cold. For winter there is a warm patch. The temprature then drops for around four days in the middle. It increases towards the end. In the min column below zero tempratures came in patches apart from one.*

(R706) *From the 18th to the 24th was the longest period with out rain. From the 8th to the 18 was the longest period with rain almost non-stop, with 26.6 millimetres. So I think that the best time to have an outdoor event in July would be from sometime between 18th and 24th. It rained 15 days and didn't rain 16 days for the month.*

Multistructural Responses

The multistructural responses discuss more than one feature of data when describing the variation and usually combine elements identified for discussing extremes with elements identified for discussing interior values. Only sixteen responses were coded at this level, all describing the temperature data and just one from a Grade 7 student. While the quality of the description of the extreme values does not vary much in these responses, the quality of the description of the interior values does. Extreme values are usually discussed as maximum and/or minimum and in some cases an actual range is given. When discussing the interior values some responses give an overview while others discuss specific data values.

Typical of those responses giving an overview of the interior values, while also mentioning the extreme values, are R909 and R910. The first gives the statistically unsophisticated *rise and fall* overview of the interior values, along with the maximum and minimum. The second states what the temperature will get down to (i.e., the minimum) but then describes each of the two temperature data sets by giving *seems to follow a bit of a pattern* as an overview of the interior minimum temperatures and *stays pretty much constant* as an overview of the interior maximum temperatures.

(R909*) In my month I can see that the highest temp was 15 degrees Celsius. The lowest temp record was -6.7 degrees Celsius. I believe this would be a bad month to hold the festivle because it is too cold. The temp pattern seems to rise and fall throughout the month.*

(R910) *The minimum temperature seems to follow a bit of a pattern. The temp. gets down to -9 degrees Celsius. The maximum temperature stays pretty much constant, it isn't affected much by the really cold minimum temperatures.*

Only five responses could be considered to have gone into more detail about the interior values while also mentioning the extreme values. Two such responses are presented here. R1108 does this by discussing, in a statistically unsophisticated way, the patches of warmer or colder weather in more detail sequentially through the month. R1109 attempts to consider a relationship between the two variables Minimum Temperature and Maximum Temperature.

(R1108) *This month is in the middle of the summer, so most of the temperatures are in the late 20's earlie 30's. The lowest max temperature is 20.4 degrees at the start of the month. The highest max temperature is the second last day of the month, temperature is 31.3 degrees. There seems to be groups of high min temperatures of 10 degrees plus, 4 or 5 high ones and then 1 or 2 low min temperatures, Where as with the max temperatures the temperature builds up for example, 22 degrees, 26 degrees, 26 degrees, 27 degrees, 29 degrees, 30 degrees, 24 degrees and then suddenly drops 5 degrees or 6 degrees.*

(R1109) *October - 00 would be a good month to hold the Youth Fest, because the max. deg of temperature varies between 11.3 and 28.3, and the min deg of temperature lies between -3.1 and 14.3. The max. temperature is high at the beginning of the October month, it slowly rises then gradually drops mid October, at this period the min degree temperature is around it's best, again the max. temp rises and drops towards the end of the month. At the same time, the min-temp rises when the max temp is remaining constant (20/21). Therefore, if the Youth fest is to be held in October, 00, it should be on a day that is included in the constant temperature pattern.*

Relational Responses

The relational responses attempt to tie together the extreme and interior values and suggest immature notions of deviations in the data values. R911 considers the day to day deviation for one 24 hour period, while R1110 considers the day to day deviation on a couple of the days and R707 discusses what appear to be, but are not obviously, 'averaged' deviations from day to day.

(R911) *The highest temperature is 20.3 degrees Celsius and the lowest temperature is -6.3 degrees Celsius. The maximum degrees Celsius ranges 13.2 degrees Celsius. The minimum temp ranges 15.7 degrees Celsius. I don't think this month would be good for the youth event as it is to cold. The temperature jumps quite a bit in places. One day the min temp was -0.9 and the next it was 7 degrees Celsius.*

(R1110) *I don't think that Jul 98 would be a very convenient time to hold the youth festival because although the weather is reasonably steady it is often very unpredictable. For example on the 4th of July the temperatures rose almost 6 degrees over night yet only a few days later on the 7th it dropped another 5.8 degrees over night. There is no real pattern here like I said the weather seems to be very unpredictable. The information shows that it is a month with moderate cool weather. The average range of the temperatures between the min and max degrees for a certain day is on the 20th when the range is 15.9 degrees Celsius and on the 28th when the range is a mere 1.1 degrees Celsius. The minimum degrees for any day is on the 3rd with -5.9 degrees Celsius.*

(R707) *I think that my month would be unsuitable to hold "Youth Fest" because the high temperatures are on average around 2 or 3 degrees different everyday. The same happened with the Minimum temperatures. They were also very cold with -3, -5 degree.*

5. DISCUSSION

This study focused on refining the Reading and Shaughnessy (2004) hierarchy based on responses from weather-related inference tasks. SOLO was used as a framework to support the refined hierarchy and two cycles of levels of cognitive growth were identified. While the Reading and Shaughnessy hierarchy was useful as a starting point, it was not detailed enough to accommodate the range of responses that were given by students. The students in the present study were engaged in a different task, involving inference from data with real variation rather than a sampling task in a probability context. Also, there was a richness in the contexts from which responses were collected, both before and after the group work and the teaching episodes. This section first addresses the three research questions proposed earlier, in the light of the results. Following that, the newly developed hierarchy is compared to hierarchies developed by other researchers and finally some limitations of the study are considered.

5.1. DESCRIPTION OF VARIATION HIERARCHY

The three research questions are now addressed. First, the specific refinements used to produce the refined hierarchy are outlined. Next, it will be argued that SOLO provides a suitable explanation for this hierarchy. Finally, the notion of two cycles of levels identified within one SOLO mode, as has been found by other researchers, is confirmed for these responses.

Refinement of the Reading and Shaughnessy (2004) hierarchy

The first research question asked whether the hierarchy developed for analyzing students' descriptions of variation in a sampling situation (Reading & Shaughnessy, 2004) was also applicable for coding responses with data descriptions given when making inferences from weather-related data, in which there is natural variation. Although the Reading and Shaughnessy (2004) hierarchy proved useful as a foundation for the coding, a more detailed structure was needed to account for the array of responses given by students. Table 9 links each of the two groupings of the hierarchy proposed by the analysis in this study to the original Reading and Shaughnessy levels on which they were based. The three levels of the first grouping, based on qualitative feature were considered by the researchers to be less statistically sophisticated versions of the D1 - *Extreme or Middle Values*, and D2 - *Extreme and Middle Values* of the Reading and Shaughnessy hierarchy. The responses described in this qualitative grouping help to give an insight into early considerations of features of variation that later develop to become the more easily recognizable numeric descriptions of variation. The three levels in the second grouping, based on quantitative features of variation, more closely align with the D1 and D2 levels of the Reading and Shaughnessy hierarchy.

Table 9. Refined hierarchy linked to the Reading and Shaughnessy (2004) hierarchy

Refinement for Description of Variation Hierarchy	**Link to Reading &Shaughnessy (2004) Hierarchy**
Qualitative Responses	Expressed in words only, with no descriptions of numeric features of variation
unistructural - one qualitative feature of variation	Like the D1 responses, *Extreme Values or Middle Values*
multistructural - more than one qualitative feature of variation	Like the D2 responses, both *Extreme Values and Middle Values*
relational - link qualitative features of variation	Links the *Extreme Values and Middle Values* features of variation
Quantitative Responses	Expressed with numeric features of variation
unistructural - one quantitative feature of variation	Equivalent to the D1 responses, *Extreme Values or Middle Values*
multistructural - more than one quantitative feature of variation	Equivalent to the D2 responses, both *Extreme Values and Middle Values*
relational - link quantitative features of variation	Links the *Extreme Values and Middle Values* features of variation, may suggest notion of deviation and hence be heading towards a D3 response.

The responses in the qualitative grouping are considered to be less statistically sophisticated than the responses in the quantitative grouping. Although students' qualitative descriptions show that they have been able to notice and acknowledge variation, they have not been able to apply a measure to their description. It should be noted that the use of the term *Middle Values* in the Reading and Shaughnessy hierarchy, meant to refer to the values not occurring at the extremes, was being misinterpreted by users of the hierarchy as referring to measures of central tendency. To avoid further confusion the terminology was changed from *Middle Values*, as used by Reading and Shaughnessy, to *Interior Values* in the refined hierarchy. The term *Middle* has still been used in Table 9, consistent with the Reading and Shaughnessy hierarchy but the term *Interior* is used in later descriptions of the refined hierarchy. The expression 'Like' is used in the explanations for the qualitative responses because these responses were describing the same sort of features as described in the Reading and Shaughnessy D1 and D2 levels but not in the same way, i.e., they did not contain the numerically described features of variation that D1 and D2 contained. The expression 'Equivalent' is used for the quantitative responses because these responses included features of variation described in the same numeric fashion as those in the D1 and D2 levels.

No responses were found in the present study that specifically discussed deviations from an anchor, central or non-central, and hence could be considered as equivalent to those identified by Reading and Shaughnessy as D3 - *Discuss Deviations from an Anchor* or D4 - *Discuss Deviations from a Central Anchor*. However, there were two responses in the present study, R1110 and R707 at the relational level, which may be considered transitional to being coded as D3 because of the attempt to describe the deviations.

Thus, in response to the first research question, it was possible to refine the Reading and Shaughnessy (2004) hierarchy by identifying responses equivalent to those in the D1 and D2 levels and by also identifying responses that were structurally similar to D1 and D2 responses but expressed in the less statistically mature qualitative form. Additional research is needed with more statistically sophisticated responses that those given in the present study to be able to refine the D3 and D4 levels, where deviations become the focus of the discussion.

Other researchers, too, have reported finding similar approaches to dealing with variation as those identified here. For example, delMas and Liu (2003) investigated students' formation of ideas when they were first learning about factors that affect standard deviation. Of interest are strategies they

identified students using when attempting to move bars in a graph to produce maximum or minimum standard deviation. One strategy, 'equally spread out', focusing on equal separation of the bars in the graph, is similar to the descriptions in the *Interior Values* focused qualitative unistructural responses identified in the present study, while the 'far-away' strategy, focusing on getting the bars as far away from each other as possible, is similar to the variation descriptions focusing on *Extreme Values*.

SOLO as a theoretical framework for the refined hierarchy

Having established that this hierarchy is suitable, the second research question asked whether SOLO could offer a broad framework for explaining the hierarchy. Discussion now focuses on explaining how the taxonomy was used to explain this cognitive growth as a distinct cycle of unistructural (U), multistructural (M) and relational (R) levels (see section 2.1 and Pegg, 2003, p. 243). Table 10 summarizes the application of the SOLO Framework to the six levels. In the qualitative responses, the first three levels now labeled as the first cycle, identification of the element of interest as 'a feature of the variation of the data described qualitatively' allows the three levels within that category, to be explained as unistructural, multistructural and relational. The unistructural (U1) responses contain one such element, the multistructural (M1) responses contain more than one such element and the relational (R1) responses link these elements. This cycle has some qualitative descriptions that are more *Sequential* in nature while others are more *Limiting*. One key to better defining what is happening in this first cycle might be to look to other research that identifies intuitive notions, such as that by Makar and Confrey (2003) who found that pre-service teachers were using 'informal' terms when comparing dotplots but in the process were discussing non-simplistic concepts. Responses that suggested consideration of clustering, as opposed to modal clumping, and the terms used by these prospective teachers may help to unravel the often-unclear terminology used by younger students and add to the definition of levels in this cycle.

Table 10. Refined description of variation hierarchy

First Cycle Qualitative Responses	**element - qualitative feature of variation of data**
U1 - unistructural - one qualitative feature of variation	magnitude related - in an absolute sense to give indication of size of change, e.g*., pretty much consistent* or arrangement related - in a relative sense to give position, e.g., *spread out pretty evenly*
M1 - multistructural - more than one qualitative feature of variation	limiting related - set limits on the data values, e.g., *doesn't get too hot* and/or sequential related - deal with data item by item, e.g., *lots of dry days then a couple of wet days then a lot of dry days again*
R1 - relational - link qualitative features of variation	link the general limit with the discussion of blocks sequentially, e.g., *seems to fall pretty regularly but the amounts are not too much .. main pattern seems to be a short spell of dry days (3-5days) and then 1 or 2 wet days but rain is pretty light and not a large amount falls…*
Second Cycle Quantitative Responses	**element - quantitative feature of variation of data**
U2 - unistructural - one quantitative feature of variation	based on extreme values - discuss maximum, minimum, range or interior values - refer to blocks or patches of days
M2 - multistructural - more than one quantitative feature of variation	based on extreme values and/or interior values, e.g., refer to range but also to the rise and fall of temperatures throughout the month
R2 - relational - link quantitative features of variation	linking of extreme values and interior values may suggest immature notions of deviations, e.g., discussions including day-to-day deviations or 'averaged' deviations from day-to-day

In the quantitative responses, the last three levels now labeled as the second cycle, identification of the element of interest as 'a feature of the variation of the data described quantitatively' explains the three levels, unistructural (U2), multistructural (M2) and relational (R2) of this cycle. This cycle of levels includes responses that clearly deal with *Extreme Values* while others deal with *Interior Values*. The importance of investigating these notions of evaluating dispersion were also borne out by the research of Lann and Falk (2003), who evaluated strategies used by statistically naive tertiary students, as they compared sequences of data for greater variability. Some criteria for making decisions were common such as the Range and Interquartile Range, while others were more intuitive and less easy to unravel. Responses using the Range would identify as having an *Extreme Values* focus, in the refined hierarchy, while those using the Interquartile Range would be classified as *Interior Values*. Lann and Falk (2003) also attempted to analyze the justifications given for selected responses but found that the analysis of these explanations was not such an easy task. Results from their, yet to be investigated, considerable number of 'no definite diagnosis' responses, may also add to the story in the second cycle but is more likely to assist in unraveling the mystery of what students really mean when they give responses such as those in the first cycle.

Two Cycles of SOLO Levels identified

Having established that SOLO proved useful as a suitable framework, the third research question asked whether cycles of levels can be identified within the SOLO modes? Two distinct cycles of the unistructural-multistructural-relational levels have been identified. Both these cycles are part of the concrete symbolic mode (Pegg, 2003, p.242) where a person thinks through the use of the symbol systems, both language and numeric, as used by the literate. Pegg (2003, p. 245) provides a useful diagrammatic representation of the link between coexisting cycles of levels within the concrete symbolic mode. As pointed out earlier the existence of more than one cycle of levels of cognitive development within one SOLO mode of cognition has now been observed by other researchers and so it is not unexpected that two cycles of levels would be observed in this study. Of particular interest with the first cycle in the present study is the strong emphasis on visual elements in the descriptions of variation. This would be expected because, as Pegg (2003, p. 244) points out, this first cycle in the concrete symbolic mode provides an interface to the less cognitively developed ikonic mode of operation, where actions are internalized as images. In fact, some responses demonstrated that students revert to the ikonic mode, based on personal experience, such as their own knowledge of festivals and the town's weather, when trying to justify their evaluation of the suitability of the month for the event.

The nature of the responses as described in the refined Description of Variation Hierarchy, within each of the two cycles, demonstrates a developmental cognitive progression from the first to second cycle. Those responses at the first multistructural level specifically coded as *Limiting* appear to be precursors to the *Extreme Value*s responses at the second unistructural level, while those coded as *Sequential* appear to be precursors to the *Interior Values*. As the terminology used by students progresses through the levels of the two cycles it appears as if the students are adjusting the focusing lens on a microscope. The higher the level of response achieved the finer the detail provided about the variation that exists. Even finer detail is expected to unfold in future research during analysis of responses to other tasks and from more advanced students.

5.2. COMPARISON TO OTHER DEVELOPMENTAL HIERARCHIES

The Description of Variation Hierarchy as refined by this study, see Table 10, provides a greater depth of explanation of the focus of student responses on variation than the previously developed hierarchies, by Mooney (2002), see Table 1, and Watson et al. (2003), see Table 2. This has been achieved by identifying cycles of levels, unistructural, multistructural and relational, within the SOLO mode of cognitive growth for the less statistically sophisticated categories in both of these hierarchies. Is the greater depth of explanation within the refined hierarchy consistent with references to spread in the Mooney (2000) hierarchy? Though variation is only acknowledged through references to spread in the *Organising and Reducing Data* process of Mooney's (2002) framework, similarities in descriptors

can be found with the refined hierarchy proposed in Table 10. Mooney's *Transitional* responses, with their 'invented' measures are similar to the proposed first cycle of qualitative responses. However, the proposed second cycle of quantitative measures has not distinguished between invented and valid measures as Mooney's *Qualitative* and *Analytical* levels have done. There were not sufficient responses at, or above, the relational level of the quantitative responses in the present study to develop the hierarchy further at these higher levels of cognition.

Is the greater depth of explanation in the refined hierarchy consistent with the hierarchy developed by Watson et al. (2003)? The ability to describe variation is essential to demonstrating the achievement of levels of understanding developed by Watson et al (2003). The qualitative responses identified in the first cycle of the refined hierarchy in Table 10 are typical of descriptions given in responses at Level 1 - *Prerequisites for Variation* of the Watson et al. hierarchy. The quantitative responses identified as second cycle of the refined hierarchy are typical of descriptions given in responses at Level 2 - *Partial Recognition of Variation*. As previously mentioned responses at, or above, the relational level of the quantitative responses were lacking and if such responses are collected in the future they may provide cycles of levels of description of variation equating to the upper two levels of the Watson et al. Hierarchy, Level 3 - *Applications of Variation* and Level 4 - *Critical Aspects of Variation*. Thus the refined hierarchy proposed by the present study has provided a greater depth of explanation to the lower cognitive levels of both the Mooney (2002) and Watson et al. (2003) hierarchies.

5.3. INTERPRETATIONS AND LIMITATIONS

Any consideration of the findings reported above needs to take into account students' interpretations of the task and two noticeable limitations of the study relating to student motivation and the profile of the data supplied for the task. Student interpretation issues focus around the initial intent of the activity, the different approaches to the two data variables and lack of recognition of the benefit of using visual representations. The way most of the students interpreted the task, though not contrary to, was not exactly what was initially intended. The natural urge to predict the most suitable month, even before being asked to do so, suggests that for students to give more meaningful descriptions of data they need a context and a sense of purpose. In this case, the students were given a context, using rainfall and temperature data from their own town for the preceding three years, and a purpose, to decide on the suitability of a particular month for the scheduling of a Youth Festival.

The students' familiarity with weather and with their expectations of the need for suitable weather for the festival may have contributed to the differing approaches that students took to describing the variation in the data for rainfall and for temperature. Consideration of the data for just part of the month was more common when describing the rainfall data, where blocking of 'rain' and 'no rain' days was often the focus. For temperature, use of the data for all of the month was more common and extremes of temperature became the focus of the better responses. Rainfall almost took on a dichotomous nature in that interest centred on whether it 'rained or not', while temperature maintained a more continuous nature with the number of degrees being considered to be of enough importance to be discussed.

The use of a realistic context, though considered more meaningful, appears to have precluded students from recognizing an opportunity to make use of skills newly acquired in the classroom. Few students beyond Grade 7 drew a graph to help describe the data and only three students referred specifically to their graphs in their explanations. Even the inclusion of a teaching episode, to introduce a new graphing technique to the students, did not result in any noticeable increase in the use of graphs to aid the inferences. It is possible that if students had been encouraged, or actually required, to draw a graph of the data then visual cues may have assisted them to give more detailed descriptions of the variation.

Student motivation was clearly evident early on in the task, but waned as the activity progressed. Well intentioned attempts to provide a realistic context for the inference task were obviously successful to the point of creating another problem. Some students thought the Youth Alive festival

was really going to take place and Grade 7 were particularly disappointed when they found out that this was not so.

Apart from the differing student interpretations of the nature of the two variables, rainfall and temperature, the profile of the data also differed in the amount of information supplied. One set of data was given per month for rainfall (daily millimetres) and two sets of data per month (daily maximum degrees centigrade and daily minimum degrees centigrade) for temperature. The two sets of data for temperature proved more of a complication for students than had been anticipated. In many cases students dealt with this issue by using only one set of data or the other, or by combining all the data into one set with maximum temperatures and minimum temperatures together.

These various interpretation issues and limitations were not considered to detract, however, from the wealth of information contained in the responses. This was especially so given that the coding of the responses was not to be used as a quantification of the best of students' capabilities but more as an indication of what descriptions of variation are used by students as they respond to the particular weather-related inference task.

6. IMPLICATIONS

6.1. IMPLICATIONS FOR RESEARCH

Several implications for research arise from this study. First, the refining of the hierarchy has demonstrated that levels devised for description of variation in sampling task responses have proven useful as a starting point for analyzing responses in an inference task and that SOLO can provide a suitable framework for such a hierarchy. Descriptions of lower level responses in the original hierarchy have been expanded and levels have been created for responses that are less statistically sophisticated than those in the original hierarchy. The strength of this refinement of a previously developed hierarchy now needs to be tested by applying the developed cycles of levels to the coding of responses posed in statistical tasks based in other contexts. Another implication of this study is that more statistically sophisticated responses need to be analyzed to identify the structure of possible cycles of levels that may exist above the two cycles proposed. It is expected that research with more advanced students will reveal some detail of more sophisticated development. The delMas and Liu (2003) research is a clear indication of the reasonableness of this expectation. A strategy they found being used by college students to describe variation, 'far-away mean', focused on trying to get the bars of a computer display as far away as possible from the mean in order to affect the standard deviation. This is similar in approach to the descriptions given by students at the Reading and Shaughnessy (2004) D4 – *Discuss Deviation from a Central Anchor* level and indicates that refinement of the D3 and D4 levels would be warranted.

A further implication is that when designing tasks researchers need to be aware of the influence of the nature of the variable used in the task on the style of response and to try encouraging the use of graphical representation to improve the quality of descriptions of variation. Related to this is the implication that care should be taken to avoid unnecessary complication in tasks given. In this case, future use of the weather activity should only include one set of data for a particular variable, e.g., the more relevant Maximum Temperature for the temperature segment of the activity. This would remove the complication, unnecessary to this particular investigation, of having to deal with two sets of data for the one variable. A final implication is that consideration in future research should also be given to the role played by measures of central tendency, such as the mean, when describing variation.

6.2. IMPLICATIONS FOR TEACHING AND ASSESSMENT

From a teaching perspective, student responses to the weather activity demonstrate that when considering data 'in context' students may rely too much on their experience of the context itself and not enough on information provided by the data. This then influences the way students describe the variation of the data, and ultimately any predictions made. It is also evident that the nature of the data

in the variable influences the way that students react to data. Students focused on varying amounts of the information depending on and treated them differently depending on whether the temperature or rainfall variable was being discussed. In addition, teachers should consider encouraging more use of graphical representation when students are engaged in activities that involve description of variation. The terminology used by students is important and there is a need to encourage students to work from their own terminology and descriptions to what is required in more statistically sophisticated discussions. Such a necessity has also been flagged by Makar and Confrey (2003). Finally, the description of responses at the various levels can be used to help teachers make sense of the unsophisticated language and reasoning of students during classroom activities.

From an assessment perspective, the hierarchy developed in this study could provide a rubric to assess the level of cognitive growth at which students are operating in terms of their description of variation, a very basic statistical concept. Such descriptions are essential if students are to be able to indicate their appreciation of existing variation and communicate such information in a statistically sophisticated manner to a wider audience. As such a hierarchy is further developed teachers should be encouraged to use it to code responses to a variety of statistical tasks, so that they will be better informed as to how students are describing the variation as part of their reasoning about variability and patterns in data.

ACKNOWLEDGEMENTS

This paper reports on research that formed the basis of the presentation *Student Perceptions of Variation in a Real World Context* at the Third International Statistical Reasoning, Thinking and Literacy Research Forum in Lincoln, Nebraska from July 23 to 28 in 2003. The author acknowledges the helpful comments made by reviewers and editors who have contributed to the revision of this paper.

REFERENCES

Bakker, A. (2003). Reasoning about shape as a pattern in variability. In C. Lee (Ed.) *Proceedings of the Third International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-3)*. [CDROM, with video segments] Mount Pleasant, Michigan: Central Michigan University.

Ben-Zvi, D. (2003). The emergence of reasoning about variability in comparing distributions: A case study of two seventh grade students. In C. Lee (Ed.) *Proceedings of the Third International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-3)*. [CDROM, with video segments] Mount Pleasant, Michigan: Central Michigan University.

Biggs, J., & Collis, K. (1991). Multimodal learning and the quality of intelligent behaviour. In H. Rowe (Ed.), *Intelligence, Reconceptualization and Measurement* (pp. 57–76). New Jersey: Laurence Erlbaum Assoc.

DelMas, R.C., & Liu, Y. (2003). Exploring students' understanding of statistical variation. In C. Lee (Ed.) *Proceedings of the Third International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-3)*. [CDROM, with video segments] Mount Pleasant, Michigan: Central Michigan University.

Jones, G. A., Langrall, C. W., Thornton, C. A., & Mogill, A. T. (1997). A framework for assessing and nurturing young children's thinking in probability. *Educational Studies in Mathematics*, *32*, 101–125.

Jones, G. A., Mooney, E. S., Langrall, C. W., & Thornton, C. A. (2002). Students' individual and collective statistical thinking. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics: Developing a Statistically Literate Society*, *Cape Town, South Africa*. [CDROM] Voorburg, The Netherlands: International Statistical Institute.

Jones, G. A., Thornton, C. A., Langrall, C. W., Mooney, E. S., Perry, B., & Putt, I. A. (2000). A framework for characterizing children's statistical thinking. *Mathematical Thinking and Learning, 2*, 269–307.

Langrall, C. W. & Mooney, E. S. (2002). The development of a framework characterizing middle school students' statistical thinking. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics: Developing a Statistically Literate Society*, *Cape Town, South Africa*. [CDROM] Voorburg, The Netherlands: International Statistical Institute.

Lann, A., & Falk, R. (2003). What are the clues for intuitive assessment of variability? In C. Lee (Ed.) *Proceedings of the Third International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-3)*. [CDROM, with video segments] Mount Pleasant, Michigan: Central Michigan University.

Makar, K., & Confrey, J. (2003). Chunks, clumps, and spread out. In C. Lee (Ed.) *Proceedings of the Third International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-3)*. [CDROM, with video segments] Mount Pleasant, Michigan: Central Michigan University.

Mooney, E. S. (2002). A framework for characterizing middle school students' statistical thinking. *Mathematical Thinking and Learning, 4*(1), 23–63.

Pegg, J. (2003). Assessment in mathematics: A developmental approach. In J. M. Royer (Ed.), *Advances in Mathematical Cognition* (pp. 227–259). Greenwich, CT: Information Age Publishing.

Reading, C. (1998). Reactions to data: Students' understanding of data interpretation. In L. Pereira-Mendoza, L. Kea, T. Kee & W-K Wong (Eds.), *Proceedings of the Fifth International Conference on Teaching Statistics* (Vol. 3, pp. 1427–1433). Voorburg, The Netherlands: International Statistical Institute.

Reading, C., & Lawrie, L. (2004). Using SOLO to analyse group responses. In M. J. Hoines & A. B. Fuglestad (Eds.), *Proceedings of the 28th International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 193–200). Bergen, Norway: Bergen University College.

Reading, C., & Pegg, J. (1996). Exploring understanding of data reduction. In A. Gutierrez (Ed.), *Proceedings of the 20th International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 187–195). Valencia, Spain: University of Valencia.

Reading, C., & Shaughnessy, J. M. (2000). Student perceptions of variation in a sampling situation. In T. Nakahar & M. Koyama (Eds.), *Proceedings of the 24th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 89–96). Hiroshima: Hiroshima University.

Reading, C., & Shaughnessy, M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.) *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 201–226). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Shaughnessy, M., & Ciancetta, M. (2002). Students' understanding of variability in a probability environment. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics: Developing a Statistically Literate Society*, *Cape Town, South Africa*. [CDROM] Voorburg, The Netherlands: International Statistical Institute.

Shaughnessy, M., Watson, J., Moritz, J., & Reading, C. (1999). School mathematics students' acknowledgement of statistical variation. In C.Maher (Chair), *There's more to life than centers*, Procession Research Symposium, 77th Annual National Council of Teachers of Mathematics Conference, San Franciso, CA.

Torok, R., & Watson, J. (2000). Development of the concept of statistical variation: an exploratory study. *Mathematics Education Research Journal*, *12*(2), 147–169.

Watson, J. M., & Kelly, B.A. (2003). Developing intuitions about variation: The weather. In C. Lee (Ed.) *Proceedings of the Third International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-3)*. [CDROM, with video segments] Mount Pleasant, Michigan: Central Michigan University.

Watson, J. M., Kelly, B. A., Callingham, R. A., & Shaughnessy, J. M. (2003). The measurement of school students' understanding of statistical variation. *International Journal of Mathematical Education in Science and Technology*, *34*(1), 1–29.

Wild, C., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223–262.

CHRIS READING
SiMERR National Centre
Education Building
University of New England
NSW 2351
Australia

APPENDIX – SAMPLE WEATHER ACTIVITY DATA HANDOUTS

Jan-98	Jul-99	Jul-98	Jul-98	Oct-00	Oct-00
Millimetres	Millimetres	Max deg C	Min deg C	Max deg C	Min deg C
0.0	21.8	8.8	3.5	22.5	7.9
0.0	0.8	12.9	0.4	19.6	-1.9
0.0	0.0	14.1	-5.9	22.0	-2.8
4.8	0.0	12.6	3.0	23.6	-3.1
2.2	0.0	18.5	4.0	24.4	3.1
9.2	0.0	18.3	2.8	25.4	3.0
0.0	0.0	15.8	10.0	26.5	3.0
0.0	0.2	10.8	8.1	27.5	4.6
0.0	0.4	8.2	4.9	28.3	9.4
11.2	6.0	8.5	-1.2	21.7	14.3
0.0	0.8	12.9	-2.8	19.4	2.3
2.2	0.0	14.0	0.2	20.1	8.0
0.0	0.8	12.0	5.1	20.0	4.2
0.0	1.0	10.0	3.5	11.3	10.1
2.4	16.2	9.7	-1.4	17.5	2.4
0.0	0.8	12.5	-2.1	19.1	1.5
0.0	0.4	16.1	-0.2	19.2	4.6
0.0	0.0	12.6	7.1	18.4	9.4
0.0	0.0	16.7	6.4	15.9	10.7
0.0	0.0	19.0	3.1	20.6	6.9
0.6	0.0	12.4	9.9	21.9	7.7
4.8	0.0	11.4	4.1	22.3	10.8
0.0	0.0	11.6	-4.2	22.0	10.4
0.0	0.0	13.2	-0.7	21.5	11.4
0.0	0.6	12.0	5.7	20.4	11.0
8.8	0.2	13.0	6.8	21.1	4.1
0.0	1.2	16.7	9.0	20.7	4.1
0.0	0.8	11.1	10.0	19.6	1.9
0.0	0.0	6.6	0.8	21.1	5.7
0.0	0.0	5.4	0.6	15.5	10.1
0.0	0.0	4.9	0.4	15.2	9.9
Rainfall Data (January 1998) used to produce response R905	Rainfall Data (July 1999) used to produce response R706	Temperature Data (July 1998) used to produce response R1110		Temperature Data (October 2000) used to produce response R1109	

FORTHCOMING IASE CONFERENCES

1. THE 2005 SESSION OF THE INTERNATIONAL STATISTICAL INSTITUTE, ISI-55
Sydney, Australia, April 5-12, 2005

Don't miss the 2005 Session of the International Statistical Institute (ISI) in Sydney, Australia. This meeting is open to all those interested in statistical matters, particularly members of the ISI and its Sections. Those who want to stay in touch with the latest research and practice developments in the field will find ISI-55 of great relevance and should register now.

The ISI 2005 registration brochure (Bulletin No 2) has been released with new details of the scientific program, registration, social events and tours, and can be found at www.tourhosts.com.au/isi2005. The scientific program is wide ranging and includes some world-class speakers. The International Association for Statistical Education will sponsor over 10 sessions on various aspects of statistics education research and practice. For a list of Invited Paper Meetings, or if you are interested in preparing a Contributed paper for the Session, see the Session website listed above for details and submission instructions.

The scientific program is supplemented with tutorials and short courses. Satellite meetings, before and after ISI 2005, will be held at interesting locations such as Cairns, Auckland and Wellington in New Zealand, and Noumea in New Caledonia. As part of ISI-55 itself, there will also be special theme days on topics of importance to the statistical community (finance and statistics, environmental statistics and genomics).

The Social Program will be a highlight of the Session and has been designed to provide participants with an opportunity to relax and maximise networking opportunities. Sydney itself is an exciting and cosmopolitan city located on one of the largest and most beautiful harbours in the world.

For more details on the 2005 ISI Session see www.tourhosts.com.au/isi2005 or email the conference managers on isi2005@tourhosts.com.au

1.1. IASE ACTIVITIES AT THE ISI-55

Chris Wild (c.wild@auckland.ac.nz) is the IASE representative at the ISI Programme Co-ordinating Committee for ISI-55th Session. The sessions approved for ISI 55 in Sydney that were sponsored or co-sponsored by IASE and their organizers are as follows:

IPM 45 *Reasoning about Variation*. Christine Reading, creading@metz.une.edu.au

IPM 46 *The use of Simulation in Statistics Education*. Andrej Blejec, andrej.blejec@nib.si

IPM 47 *Teaching Statistics Online*. Larry Weldon, weldon@sfu.ca

IPM 48 *Statistics for Life: What are the Statistical Ideas or Skills that Matter most and why?* Chris Wild, c.wild@stat.auckland.ac.nz

IPM 49 *Research in Statistical Education*. Kay Lipson, klipson@swin.edu.au / Maria Ottaviani, Mariagabriella.ottaviani@uniroma1.it

IPM 50 *Quality Assurance in Statistics Education*. Matthew Regan, m.regan@auckland.ac.nz

IPM 51 *Promotion of Statistical Literacy among Students*. Pilar Guzman, pilar.guzman@uam.es

IPM 52 *Using History of Statistics to Enhance the Teaching of Statistics*. Carol J. Blumberg, cblumberg@winona.edu

IPM 63 *Educating the Media on how best to Report Statistics*. Jacob Ryten, rytenjacob@msn.com

IPM 81 *Ethical Standards in Statistics Education*. Mary A. Gray, mgray@american.edu

IPM 83 *Challenges in the Teaching of Survey Sampling*. Wilton de Oliveira Bussab, bussab@fgvsp.br

2. IASE SATELLITE CONFERENCE - STATISTICS EDUCATION AND THE COMMUNICATION OF STATISTICS
Sydney, Australia, April 4-5, 2005

This conference, focused on Statistics Education and the Communication of Statistics, is jointly organised by the IASE and the Victorian Branch of the Statistical Society of Australia and will immediately precede the International Statistical Institute Session in Sydney. The approach will be non-technical, suitable for both a specialist and non-specialist audience who would like to learn how to better communicate the statistical ideas which occur in their everyday and working lives. This meeting is intended to be of interest to a wide cross section of society including teachers, educational administrators, researchers in statistical education and in probabilistic reasoning and others who want to gain a better grasp of how to communicate statistics in general and who would like to broaden their knowledge of statistics applications. It should also be of interest to people concerned with interpreting sociological, economical, political, scientific or educational reports, predicting sports results, by policy makers, journalists, health professionals and others from the general population.

Over 25 abstracts were submitted by the Sept 30 deadline, showing there is a lot of interest in this topic. If you are going to ISI and are interested in issues related to the communication of statistics or statistics education, you should make a special effort to attend this meeting.

For details see the web site: www.stat.auckland.ac.nz/~iase/conferences.php?show=iase2005 or contact the joint chairs, Brian Phillips, bphillips@swin.edu.au and Kay Lipson, klipson@swin.edu.au

3. SRTL-4 THE FOURTH INTERNATIONAL RESEARCH FORUM ON STATISTICAL REASONING, THINKING AND LITERACY
Auckland, New Zealand, July 2-7, 2005

The Fourth International Research Forum on Statistical Reasoning, Thinking, and Literacy, is to be hosted by the Department of Statistics, The University of Auckland, New Zealand, July 2–7, 2005. This gathering offers an opportunity for a small, interdisciplinary group of researchers from around the world to meet for a few days to share their work, discuss important issues, and initiate collaborative projects. Having emerged from the three previous forums, the topic and focus of SRTL-4 will be Reasoning about Distribution. The Forum is co-chaired by Dani Ben-Zvi (University of Haifa, Israel) and Joan Garfield (University of Minnesota, USA), co-organized by Maxine Pfannkuch and Chris Wild (The University of Auckland, New Zealand), and planned by a prestigious international advisory committee.

Based on the SRTL tradition, we plan to keep the number of participants small to facilitate a working research forum. There are three possible roles for participants in this Forum. The first role is to present current research on reasoning about distribution, the second is to discuss and react to research presentations, while the third is to be a small group moderator, which is ideal for doctoral students who are not yet ready to present research but want to participate. Participants will be strongly encouraged to use videotape and written transcripts of students in classroom and interview settings to provide illustrations of what the researchers are learning about how students reason about distribution. As with the previous SRTL Research Forums, we encourage the participation of young promising scholars. One outcome of the Forum will be a publication summarizing the work presented, discussions conducted, and issues emerging from this gathering.

The SRTL-4 Research Forum organizers invite anyone interested in participating in this forum to contact them as soon as possible. The first deadline for submission of interest was June 1, 2004. More Information from Maxine Pfannkuch, m.pfannkuch@auckland.ac.nz.
Web site: www.stat.auckland.ac.nz/srtl4/

4. ICOTS-7: WORKING COOPERATIVELY IN STATISTICS EDUCATION
Salvador (Bahia), Brazil, July 2-7, 2006

The International Association for Statistical Education (IASE) and the International Statistical Institute (ISI) are organizing the Seventh International Conference on Teaching Statistics (ICOTS-7) which will be hosted by the Brazilian Statistical Association (ABE) in Salvador (Bahia), Brazil, July 2-7, 2006.

The major aim of ICOTS-7 is to provide the opportunity for people from around the world who are involved in statistics education to exchange ideas and experiences, to discuss the latest developments in teaching statistics and to expand their network of statistical educators. The conference theme emphasises the idea of *cooperation,* which is natural and beneficial for those involved in the different aspects of statistics education at all levels.

4.1. CALL FOR PAPERS

Statistics educators, statisticians, teachers and educators at large are invited to contribute to the scientific programme. Types of contribution include *invited papers, contributed papers* and *posters*. No person may author more than one Invited Paper at the conference, although the same person can be co-author of more than one paper, provided each paper is presented by a different person.

Voluntary refereeing procedures will be implemented for ICOTS7. Details of how to prepare manuscripts, the refereeing process and final submission arrangements will be announced later.

Invited Papers

Invited Paper Sessions are organized within 9 different Conference Topics 1 to 9. The list of Topic and Sessions themes, with email contact for Session Organisers is available at the ICOTS-7 web site at www.maths.otago.ac.nz/icots7, under “Scientific Programme”. Those interested in submitting an invited paper should contact the appropriate Session Organiser before December 1, 2004.

Contributed Papers

Contributed paper sessions will be arranged in a variety of areas. Those interested in submitting a contributed paper should contact either Joachim Engel (Engel_Joachim@ph-ludwigsburg.de) or Alan MacLean (alan.mclean@buseco.monash.edu.au) before September 1, 2005.

Posters

Those interested in submitting a poster should contact Celi Lopes (celilopes@uol.com.br) before February, 1, 2006.

Special Interest Group Meetings

These are meetings of Special Interest Groups of people who are interested in exchanging and discussing experiences and/or projects concerning a well-defined theme of common interest. Proposals to hold a SIG Meeting specifically oriented to reinforce Latin American statistics education cooperation in a particular theme are especially welcome. In this case the organisers may decide to hold the meeting in Portuguese and Spanish language. Individuals or groups may submit proposals to establish a Special Interest Group to Carmen Batanero at (batanero@ugr.es).

4.2. TOPICS AND TOPIC CONVENORS

Topic 1. *Working cooperatively in statistics education*. Lisbeth Cordani, lisbeth@maua.br and Mike Shaughnessy, mike@mth.pdx.edu

Topic 2. *Statistics Education at the School Level*. Dani Ben-Zvi, benzvi@univ.haifa.ac.il and Lionel Pereira, lpereira@nie.edu.sg

Topic 3. *Statistics Education at the Post Secondary Level*. Martha Aliaga, martha@amstat.org and Elisabeth Svensson, elisabeth.svensson@esi.oru.se

Topic 4. *Statistics Education/Training and the Workplace*. Pedro Silva, pedrosilva@ibge.gov.br and Pilar Martín, pilar.guzman@uam.es

Topic 5. *Statistics Education and the Wider Society*. Brian Phillips, BPhillips@groupwise.swin.edu.au and Phillips Boland, Philip.J.Boland@ucd.ie

Topic 6. *Research in Statistics Education*. Chris Reading, creading@metz.une.edu.au and Maxine Pfannkuch, pfannkuc@scitec.auckland.ac.nz

Topic 7. *Technology in Statistics Education.* Andrej Blejec, andrej.blejec@nib.si and Cliff Konold, konold@srri.umass.edu

Topic 8. *Other Determinants and Developments in Statistics Education.* Theodore Chadjipadelis, chadji@polsci.auth.gr and Beverley Carlson, bcarlson@eclac.cl

Topic 9. *An International Perspective on Statistics Education*. Delia North, delian@icon.co.za and Ana Silvia Haedo, haedo@qb.fcen.uba.ar

Topic 10. *Contributed Papers*. Joachim Engel, Engel_Joachim@ph-ludwigsburg.de and Alan McLean, alan.mclean@buseco.monash.edu.au

Topic 11. *Posters*. Celi Espasandín López, celilopes@directnet.com.br

4.3. ORGANISERS

Local Organisers

Pedro Alberto Morettin, (Chair; pam@ime.usp.br), Lisbeth K. Cordani (lisbeth@maua.br), Clélia Maria C. Toloi (clelia@ime.usp.br), Wilton de Oliveira Bussab (bussab@fgvsp.br), Pedro Silva (pedrosilva@ibge.gov.br).

IPC Executive

Carmen Batanero (Chair; batanero@ugr.es), Susan Starkings (Programme Chair; starkisa@lsbu.ac.uk), Allan Rossman and Beth Chance (Editors of Proceedings; arossman@calpoly.edu; bchance@calpoly.edu), John Harraway (Scientific Secretary: jharraway@maths.otago.ac.nz), Lisbeth Cordani (Local organisers representative; lisbeth@maua.br).

More information is available from the ICOTS-7 web site at www.maths.otago.ac.nz/icots7 or from the ICOTS IPC Chair Carmen Batanero (batanero@ugr.es), the Programme Chair Susan Starkings (starkisa@lsbu.ac.uk) and the Scientific Secretary John Harraway (jharraway@maths.otago.ac.nz).

OTHER FORTHCOMING CONFERENCES

1. ASIAN TECHNOLOGY CONFERENCE IN MATHEMATICS 2004
Singapore, December 13-17, 2004

There is little doubt that technology has made an impact on the teaching of Mathematics. The Asian Technology Conference in Mathematics on the theme *Technology in Mathematics: Engaging Learners, Empowering Teachers and Enabling Research* is hosted by the National Institute of Education, Nanyang Technological University of Singapore (December, 13-17, 2004). In this conference, we shall go beyond justifying the use of technology in Mathematics to discuss and examine the best practices of applying technology in the teaching and learning of Mathematics and in Mathematics research. In particular, the conference will focus on how technology can be exploited to enrich and enhance Mathematics learning, teaching and research at all levels. The conference will cover a broad range of topics on the application and use of technology in Mathematics research and teaching. More information from Wei-Chi YANG, wyang@radford.edu, and Tilak de Alwis, tdealwis@selu.edu.
Web site: www.atcminc.com/mConferences/ATCM04/

2. CONGRESS OF THE EUROPEAN SOCIETY FOR RESEARCH IN MATHEMATICS EDUCATION, CERME-4
Sant Feliu de Guíxols, Spain, February 17-21, 2005

The Fourth Congress of the European Society for Research in Mathematics Education (ERME) will be held in Sant Feliu de Guíxols, Spain, from 17 to 21 February, 2005. The conference will focus mainly on work in Thematic Groups in a style similar to that developed in previous conferences. Details of the groups from CERME3 can be found on the website www.dm.unipi.it/~didattica/CERME3/second.html. Many of the previous groups will continue work, and we expect a few new groups. Of special interest is Group 5 Stochastic Thinking (Research on probabilistic and statistical thinking; Subjects include: stochastic thinking, including probability, statistics and the interface between these domains), organized by Dave Pratt (e-mail: dave.pratt@warwick.ac.uk). CERME4 will also include plenary activities and poster presentations. More information from the Chair of the CERME4 Programme Committee: Barbara Jaworski, barbara.jaworski@hia.no. Web site: cerme4.crm.es.

3. UNITED STATES CONFERENCE ON TEACHING STATISTICS, USCOTS
Columbus, OH, USA, May 19-21, 2005

The first United States Conference on Teaching Statistics (USCOTS) will be held on May 19-21, 2005 at the Ohio State University in Columbus, Ohio, hosted by CAUSE, the Consortium for the Advancement of Undergraduate Statistics Education. USCOTS is an active, hands-on working conference for teachers of Statistics at the undergraduate level, in any discipline or type of institution, including high school teachers of AP Statistics. USCOTS will feature spotlight sessions, plenary talks, and working breakout sessions in three major areas: curriculum, pedagogy, and research. Lots of good resources for each of these areas will be provided in a fun and active atmosphere, where everyone will be invited to be involved. The theme of the 2005 USCOTS is “Building Connections for Undergraduate Statistics Teaching” and will focus on ways that we can share teaching ideas, develop working relationships, and identify areas for future collaborations and projects at our own institutions. USCOTS is partially funded by The Ohio State University Department of Statistics and its College of Mathematical and Physical Sciences. For more information about USCOTS, please contact Deborah Rumsey, USCOTS program chair at rumsey@stat.ohio-state.edu. Web site: www.causeweb.org/uscots/

4. INTERNATIONAL CONFERENCE ON MATHEMATICAL MODELLING AND APPLICATIONS, ICTMA 12
London, United Kingdom, July 10-14, 2005

Mathematical modelling and applications, the transition between real world problems and mathematical representations of such problems, is an enduring and important feature of industry, business and commerce. Teaching mathematical modelling, through tasks, projects investigations and applications embedded in courses and of mathematics itself through applications helps learners to understand relationships between real world problems and mathematical models. The 12th International Conference on Mathematical Modelling and Applications (ICTMA12) will be hosted by the School of Engineering and Mathematical Sciences City of London, Sir John Cass Business School, London, UK. This conference brings together international experts in a variety of fields and from different sectors to consider: modelling in business and industry, evaluating effectiveness, pedagogic issues for learning, applicability at different levels, research: education and practice, innovative practices and transitions to expert practice. More information from Chris Haines, ictma12@city.ac.uk.
Web site: www.city.ac.uk/conted/reseach/ictma12/index.htm

5. PSYCHOLOGY OF MATHEMATICS EDUCATION, PME-29
Melbourne, Australia, July 10-15, 2005

The PME29 conference will be held on July 10-15, 2005 in Melbourne, Australia. More information from Helen Chick, h.chick@unimelb.edu.au.
Web site: staff.edfac.unimelb.edu.au/~chick/PME29/

6. THE 25TH EUROPEAN MEETING OF STATISTICIANS, EMS 2005
Oslo, Norway, July 24-28, 2005

The meeting will cover all areas of methodological, applied and computational statistics, probability theory and applied probability. There will be 8 special lecturers, 23 ordinary invited sessions and one invited discussion session. The scientific programme is broad, with ample space for applications; invited speakers and sessions have been chosen with the specific aim to appeal to a wide audience and will bridge between theory and practice, inference and stochastic models. The meeting is organised jointly by the University of Oslo and the Norwegian Computing Center. For further informations contact email to ems2005@nr.no or visit the web site: www.ems2005.no

7. BEYOND THE FORMULA IX
Rochester, NY, USA, August 4-5, 2005

"Constantly Improving the Teaching of Introductory Statistics" is the motto of the conference organized by Monroe Community College, Rochester, NY. Beyond The Formula is an annual two-day, summer conference designed to promote excellence in teaching introductory statistics. Whether participants come from the high school, two-year or four-year college setting, they can expect to hear speakers who will provide them with new ideas and techniques that will make their classrooms more effective statistics learning centers. In organizing BTF conferences four major areas of teaching concern have been identified: the curriculum, the techniques and methodologies, the ever-changing technology, and real world applications. Each year one of these topic areas is chosen to serve as a common thread to draw together all the presentations - with 20 to 25 large-group, small-group and workshop sessions, all of the four topic areas find their way into each conference. Come with us each summer on a trip Beyond The Formula to see how exciting, thought provoking and inspiring the

teaching of statistics can really be. For more information contact Robert Johnson rr.bs.johnson@juno.com. Web site: www.monroecc.edu/go/beyondtheformula/

8. JOINT STATISTICAL MEETINGS, JSM 2005
Minneapolis, MN, USA, August 7-11, 2005

JSM (the Joint Statistical Meetings) is the largest gathering of statisticians held in North America. It is held jointly with the American Statistical Association, the International Biometric Society (ENAR and WNAR), the Institute of Mathematical Statistics, and the Statistical Society of Canada. Attended by over 4000 people, activities of the meeting include oral presentations, panel sessions, poster presentations, continuing education courses, exhibit hall (with state-of-the-art statistical products and opportunities), placement service, society and section business meetings, committee meetings, social activities, and networking opportunities. Minneapolis is the host city for JSM 2005 and offers a wide range of possibilities for sharing time with friends and colleagues. For information, contact jsm@amstat.org or phone toll-free (800) 308-8943. For more info contact email Elaine Powell, jsm@amstat.org.
Web site: www.amstat.org/meetings/jsm/2005/

STATISTICS EDUCATION RESEARCH JOURNAL REFEREES
DECEMBER 2003 – NOVEMBER 2004

Acknowledgement: We are grateful to the following IASE members and researchers who acted as referees for the Statistics Education Research Journal in the period from December 2003 to November 2004:

Richard Alldredge, USA
Paul Ayres, Australia
Arthur Bakker, The Netherlands
Gabriella Belli, USA
Nick Broers, the Netherlands
Gail Burrill, USA
Carmen Capilla, Spain
Cesar Sáenz Castro, Spain
Megan Clark, New Zealand
Jon Cryer, USA
Geoff Cumming, Australia
Bob delMas, USA
Anne Marie Dussaix, France
Jose Fernandez, Portugal
Mary Fraire, Italy
Sue Gordon, Australia
Brian Greer, USA
Jim Hammerman, USA
John Harraway, New Zealand
Ruth Heaton, USA
Hanan Innabi, United Arab Emirates
Reuben Klein, USA
Sharon Lane-Getaz, USA
Cynthia Langrall, USA
Carl Lee, USA
Yan Liu, USA
Katie Makar, USA
John McKenzie, USA
Maria Meletiou, Cyprus
Jamie Mills, USA
Gianfranco Moncecchi, Italy
John O'Donoghue, Ireland
Jenny Pange, Greece
Maria Pannone, Italy
Maria Pia Perelli, Italy
Peter Petocz. Australia
Anna Reid, Australia
Omar Rouan, Morocco
Ernesto Sanchez, Mexico
Candace Schau, USA
Gilberte Schuyten, Belgium
Norean Sharpe, USA
Mike Shaughnessy, USA
Nigel Smeeton, UK
James Tarr, USA
Dirk Tempelaar, Netherlands
Jessica Utts, USA
Ann-Lee Wang, Malaysia
Andrew Zieffler, USA