The Emergence of Reasoning about Variability in Comparing Distributions

A Case Study of Two Seventh Grade Students

Dani Ben-Zvi, University of Haifa, Faculty of Education
dbenzvi@univ.haifa.ac.il

Abstract

The purpose of the study, research question, methods, subjects age/level, what you learnt from the research).

Overview

Variability and comparing data sets stand in the heart of statistics theory and practice. "Variation is the reason why people have had to develop sophisticated statistical methods to filter out any messages in data from the surrounding noise" (Wild & Pfannkuch, 1999, p. 236). 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 (Watson & Moritz, 1999; Konold and Higgins, 2003). Thus, both variation and comparing groups deserve attention from the statistics education community.

The focus in this paper is on the emergence of beginners' reasoning about variation in a comparing groups situation during their encounters with Exploratory Data Analysis (EDA) curriculum in a technological environment. The current 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 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 numerical, tabular and graphical statistical representations. In the light of the analysis, a description of what it may mean to begin reasoning about variability in comparing distributions is proposed, and implications are drawn.

Literature Review

Research on Variation

Pfannkuch and Wild (in press) emphasize 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."

Studies of reasoning about variation includes investigations into the role of variation in graphical representation (Meletiou & Lee, 2002), comparison of data sets (Watson & Moritz, 1999; Watson, 2001; Makar & Confrey, in press), probability sample space (Shaughnessy & Ciancetta, 2002), chance, data and graphs, and sampling situations (Watson & Kelly, 2002), and variability in repeated samples (Reading & Shaughnessy, in press). Hierarchies to describe various aspects of variation and its understanding have been developed by Watson, Kelly, Callingham, and Shaughnessy (in press) and by Reading and Shaughnessy (in press) 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 (in press) discuss three different ways that teachers considered 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.

Research on Comparing Groups

Comparing groups provides the motivation and context for students to consider data as a distribution and take into account measures of variation as well as 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, in press). Watson & 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.

However, there is some evidence 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, in press), representativeness (Mokros and Russel, 1995), variability in data (e.g., Meletiou, 2002); as well as 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, in press).

Watson and Moritz (1999) observed two response levels in group comparison task 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 and Friel, 1998; Gal, Rothschild and Wagner, 1990; Hancock, Kaput & Goldsmith, 1992; Konold, Pollatsek, Well, and Gagnon, 1997; Watson & Moritz, 1999). Konold and colleagues (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.

The current study investigates the building of an intuitive foundation for reasoning about variability in meaningful and real context of comparing two data sets of equal sizes.

Context and setting

Theoretical Perspectives

Research on mathematical cognition in the last 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 development and "socialization processes" into the culture and values of "doing mathematics" (*enculturation*). Many researchers have been working on the design of 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 EDA focusing on reasoning about variability in comparing groups.

Enculturation Processes in Statistics Education

One of the ideas used in this study is that of *enculturation.* Recent learning theories in mathematics education (cf., Schoenfeld, 1992; Resnick, 1988) include the process of enculturation. Briefly stated, this process refers to entering a community or a practice and picking up their 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.

Statistical Thinking about 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. According to them, 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 (Wild & Pfannkuch, 1999, pp. 226–227). Reading and Shaughnessy (in press) suggest two additional aspects of variation that need to be considered—describing and representing. They claim that "much of the uncertainty that needs to be dealt with when thinking statistically stems from omnipresent variation and from these six aspects of variation that form an important foundation for statistical thinking."

Based on these perspectives, the following research question was used to structure the case study and the analysis of data collected: How do junior high school students begin to reason about variability in comparing groups, in the context of open-ended problem-solving situation, supported by computerized tools?

Method

This study employs a qualitative analysis method to examine seventh-grade students' statistical reasoning about variability in comparing groups in the context of a classroom investigation. Descriptions of the setting and curriculum are followed by a profile of the students, technology used, and then methods of data collection and analysis.

The Setting

The study took place in three seventh-grade classes (13-year-old girls and boys) in a progressive experimental school in Tel-Aviv. Skillful and experienced teachers, who were aware of the spirit and goals of the curriculum (described briefly later), taught the classes. They were part of the

CompuMath curriculum development and research team, which included several mathematics and statistics educators and researchers from the Weizmann Institute of Science in Israel. The CompuMath Project is a large and comprehensive mathematics curriculum for grades 7–9 (Hershkowitz, Dreyfus, Ben-Zvi, Friedlander, Hadas, Resnick, Tabach, & Schwarz, 2002) that 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.

The *Statistics Curriculum* (SC)—the data component of the CompuMath Project—was developed 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 design of the curriculum was based on the creation of small scenarios through which students can experience 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 was designed on the basis of the theoretical perspectives on learning and the expert view of statistical thinking previously described. It stresses: (a) student's active participation in organization, description, interpretation, representation, and analysis of data situations (on topics close to the students' world such as sport records, lengths of people's names in different countries, labor conflicts, car brands), with a considerable use of visual displays as analytical tools (in the spirit of Garfield, 1995, and Shaughnessy, Garfield, & Greer, 1996); and (b) incorporation of 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 2½ months, and includes student book (Ben-Zvi & Friedlander, 1997a) and teacher guide (Ben-Zvi & Ozruso, 2001).

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), the main reasons for choosing this software were:

1. Spreadsheets provide direct access that allows students to view and explore data in different forms, investigate different models that may fit the data, manipulate a line to fit a scatter plot, etc.
2. 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 directly data points 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.
3. 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 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.

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 within their regular classroom periods and included being videotaped and interviewed (after class) as well as furnishing their notebooks for analysis.

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 diagrams. 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.

Data Collection

To study the effects of the new curriculum, student behavior was analyzed using video recordings, classroom observations, interviews, and the assessment of students' notebooks and research projects. The two students—*A* and *D*—were videotaped at almost all stages (20 hours of tapes) and their notebooks were also collected.

Analysis

The analysis of the videotapes was based on interpretive microanalysis (see, for example, Meira, 1991, pp. 62–63): 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 CompuMath researchers' team and one with four researchers from the Weizmann Institute of Science, 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) attempts 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.

The *Surnames* Activity

The Surnames activity was the second data investigation activity of the SC. The students were 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 which were given to them (35 English names). Three methods were offered to compare distributions: frequency (absolute and relative frequency tables), basic central tendency and dispersion measures (e.g., mode, mean, median and range), and graphical representations (series comparison graph/double bar chart). These methods and tools were introduced to help students' in describing and interpreting the data, searching for trends and drawing conclusions on comparing the groups. The purpose of the activity was to set the stage for students to consider data as a distribution, take into account measures of variation as well as center, and notice and intuitively deal with the variability within and between distributions. The whole activity took place during approximately five 45-minute lessons.

Results: Emerging Reasoning about Variability in Comparing Distributions

The 'Story' in Brief

In this paper, I describe how *A* and *D*'s novice views slowly changed and evolved towards an expert perspective. I focus 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 'story' includes descriptions of how *A* and *D* developed an understanding of:

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

The 'Story' in Detail

In a preparatory lesson, that took place before the Surnames activity, 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 and interpreted the data, compared the groups, and composed a summary report for a shoe company. Several statistical concepts were informally introduced, such as, statistical question and hypothesis, sample, categorical and quantitative variables, absolute and relative frequency, and frequency table.

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 computer lab to work in pairs through the questions of the activity.

On What to Focus: Beginning from irrelevant and local information

After *A* and *D* added the names of their class mates to the table (a part of it presented in Table 1), they were first asked to "*Look at the table and suggest a research question about length of surnames*." The raw data was 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 to 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 learning.

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

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

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*. [Comments in block parentheses are suggested by the author, and were verified by a triangulation process. The rows are numbered to facilitate discussion in the SRTL-3 session.]

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 ("*we don't have anything interesting*") is similar to their reaction at the beginning of the first problem situation in the SC – *Olympic Records* (analyzed in detail in Ben-Zvi and 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. This finding is consistent with several other research findings (see, for example, Moschkovich et al., 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.

In sum, 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 global features of the distributions. Similarly, in the first activity of the SC, *A* and *D* were attentive to the prominence of "local deviations" in data, which also kept them from creating more global interpretations of data. Only after the following teacher's intervention were they able to start focusing on relevant information, taking into account the variability in the data.

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 (see Video Segment 1, Part I). [Full transcripts of this paper's video segments are included.]

1	*A*	[Asking the teacher] What does it mean?
2	*D*	What does it mean to "phrase a hypothesis about interesting phenomena"?
3	*A*	That there are many names beginning with 'Mc'?
4	*T*	About the length of surnames. OK?
5	*A*	What is 'interesting phenomena'?
6	*T*	Are there no interesting phenomena in the data?
7	*A*	[Cynically] It's very interesting that there is a Michael…
8	*T*	You are asked about length!

9	*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].
10	*T*	OK. You suggest that there are very short names and very long ones.
11	*A*	Do we have to compare?
12	*D*	So what's the hypothesis?
13	*T*	I don't know. First, it's a phenomenon. What do you think? Are there many long or many short?
14	*A*	There will be a lot more of the long in US.
15	*D*	More long than short.
16	*T*	OK. You have a hypothesis: In the US…
17	*A*	But what is long, and what is short?
18	*T*	That's a different question.
19	*A*	What should we write?
20	*D*	Perhaps longer than this? Or…
21	*A*	What name is considered long?
22	*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 seem 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 and Arcavi, 2001).

The teacher's initial help consisted of calling twice their attention to the investigated variable, namely, the length of a surname. Only her second trial pushed *D* to compare the surname length of two students (one from each class): "*An interesting phenomenon is that there is a 1, 2, 3, 4, 5, ... [7]-letter name here and a 4 there*". 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, "*There are very short names and very long ones*". This intervention represents a 'jump' by the teacher not reflected in student's previous comments. She then nudged them to quantify the variability in the data in a simple way, "*Are there many long or many short* [surnames]*?*"

In response to the teacher's direct question, *A* suggested that "*There will be a lot more of the long* [surnames] *in US*", echoed by D's addition, "*More long than short*". 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. Students' first attempts to describe the variability in the data by comparing long and short names raised a new question by *A*, "*what is long, and what is short?*" This concern was not resolved at this stage, and may be the beginning of an attempt to handle variability by clumping /grouping the data. The interaction with the teacher closed with her recommendation to focus on comparing groups.

In sum, the students began by considering irrelevant issues, were guided by the teacher to focus on the surnames length, picked out and compared two adjacent values, suggested a dichotomous comparison between long and short names (and possibly between the countries), and considered the definition of the borderline between long and short names.

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 (see Video segment 1, Part II).

23	*D*	Our hypothesis about interesting phenomena in the length of surnames is: In the US, surnames will be...
24	*A*	Will be longer...
25	*D*	Longer than in Israel…
26	*A*	Usually than in Israel...
27	*D*	Usually, not always, usually.
28	*D*	Let's see, we have Levkowitch here [in the Israeli class] and Cose there [in the American class] – that's different.
29	*A*	Enough, enough, come on.
30	*D*	OK, never mind.
31	*A*	So, in the US... the surnames…
32	*D*	Will be usually longer.
33	*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, i.e., surnames in the US 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*". Understanding that some surnames can behave differently even though they formulated a general 'rule', 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 US 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 (line 9). 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 including 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. Thus, using the same exploratory learning environment that they had used in algebra, 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 is 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 "*Usually, not always*" comment).

At the end of this episode they seemed very satisfied with their answer. However, it is hard to predict at this stage how fragile their current understanding is in terms of establishing abilities to acknowledge, explain, describe and deal with the variability in data in the context of this "noisy" and complex situation.

How to account for variability when comparing groups using frequency tables

After the students formulated a research question and hypothesis they learnt 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, explaining every step and overcoming technical and conceptual hurdles. The following dialogue took place when they completed the production of two univariate frequency tables (see Tables 2 and 3) and were asked to use them to compare the two groups (Video segment 2).

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%**

Table 2. 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%**

Table 3. Frequency table of surname's lengths in the American class

1 *D* [Reads the task] Use the frequency tables to compare the surnames' length in the two countries.
2 *D* Emm… They [the American surnames] are really a little bit longer. In the US there are no 2 or 3-letter names…
3 *A* Yes. And in Israel…
4 *D* … since they [the 2 or 3-letter names] are a bit short.
5 *A* The table [Table 2] starts from…
6 *D* From 2 [letters] to 8 [letters].
7 *A* The surname length is from 2 to 8.
8 *A* And in the US they're from 4 to 11.
9 *A* In other words, in the US 2 or 3-letter names are not considered at all.
10 *D* They're considered, but there are simply none.
11 *A* There are none, or there is exactly one in the whole US, something like that.
12 *A* And in Israel, names with 9, 10, and 11 letters are not considered, because there are none.
13 *D* Because they [American names] have vowels. For example, Raz, Itzik Raz [a student in their class]: Here [in Israel] it's R and Z, and there [in the US] it's R, A, and Z – three letters, did you understand?
14 *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.
15 *D* Like Levkowitch.
16 *A* So, for example, we see that names with 8 letters are 6% in Israel.
17 *D* There – they are 26%.
18 *A* In the US they are 26%.
19 *D* 20% more.
20 *A* 20% more, and it's a lot more, and…
21 *D* A lot more, interesting, lovely.
22 *D* Actually, emm… just a second…
23 *D* That's exactly all I'm saying…
24 *A* I assert that in the US there are more… the names…
25 *D* There are longer names, right.
26 *A* Longer according to the comparison between these tables. It may not be certain, but at least according to these tables.

27 *A* So, in the US 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 answer in his notebook.]

28 *A* Now, we also see here that in Israel, there are many more 4-letter names, which is considered pretty short.

29 *D* Having a 4-letter name is the coolest matter in Israel.

30 *A* So maybe because of that, there are more of those in Israel, and in the US – 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 (Tables 2 & 3). Their purpose was to find ways to justify their hypothesis that surnames in the US 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. The following diagram (Figure 1) is a visual illustration of what they were focusing on during their argumentation.

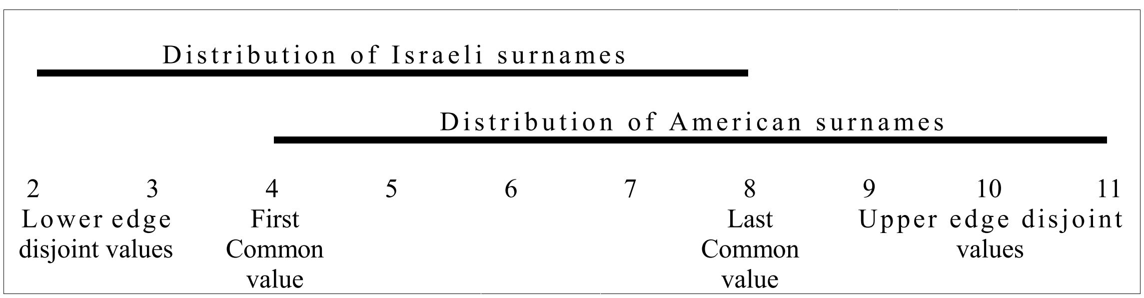

Figure 1. Illustration showing *A* and *D*'s justification structure based on disjoint and common-edge values.

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 US group) as a justification for the claim that "*They* [the American surnames] *are really a little bit longer*". They continued by noticing the different ranges of the groups; however, they did not make explicit use of them as measures of dispersion. Then *A* argued symmetrically about the right "tail" of the US distribution that is missing in Israel. This opposite symmetry between the distribution edges seems to strengthen their confidence in the claim that the US surnames are longer. The initial focus of beginners on distribution edges was found in other studies. For example, Biehler (2001) describes how novice students focused first on the "least" and the "most" while describing the variability between two distributions using box plots.

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, see Figure 1). *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 (“*Having a 4-letter name is the coolest matter in Israel*”). 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 punctuation symbols in Hebrew. He also provided an example of one Israeli surname Raz, which has three letters in English but only two in Hebrew. *A* further speculated that their sample implied the rarity of very short and very long surnames in the US and the Israeli populations respectively (e.g., “*There may be one or two all over the country*”). *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 sphere of the problem supports their statistical reasoning by providing reasonable explanations to the observed patterns in the variation.

At the end of this dialogue they wrote 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, his conclusion “*In Israel, short names are more frequent; in the USA, the long names are more frequent*” is a beginning step to modeling variability and conceptualizing the use of ‘density’ in comparing distributions.

This portion of the transcript shows the students experimenting with comparing groups in various local ways without having acquired a full conceptualization of the distribution as an entity. They used special local values (edges and neighbors to edges) as stepping stones to construct global descriptions of the groups’ differences and the variability between the distributions.

How to use center and spread measures to compare groups

In the second part of the Surnames activity the students learned basic statistical measures of center (mode, mean, and median) and spread (range), and about outliers. They used the computer to find the statistical measures of the two groups and organized them in a table (Table 4). 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 (see the following two dialogues; Video segment 3, parts I & II).

Statistical Measures	Israeli Class	US 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

Table 4. Statistical measures of the two classes. (*The median of the US group should be 7.)

1	*A*	Both the largest value and the smallest value are smaller than in the US.
2	*D*	Correct.
3	*A*	Therefore the range is also smaller.
4	*D*	Correct.
5	*D*	Ok. The range is also smaller since it's actually the range between the two values. It means that it is related to them.
6	*A*	[Cynically] Really?
7	*D*	The mean is also smaller in Israel.
8	*A*	In short [laughing], everything is smaller, and… both the mean and the median are also smaller.
9	*D*	Correct, the median too.

Using the statistical measures table (Table 4), 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, "*therefore the range is also smaller*". 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 "*everything is smaller*", meaning that all the statistical measures of the Israeli distribution are smaller than those of the US distribution. In spite of their fluent work, their actions seem to be merely procedural, missing both the meaning of measures as representative numbers, and the distinction between center and spread measures.

How to informally model variability through handling outlying values

Dealing with information in the last row of the measures table (Table 4) initiated the following dialogue about outliers (Video segment 3, part II).

10	*D*	But in the outlying values…
11	*A*	In fact here it's [different]… You expect that in Israel the outlying values will be higher than in the US, since there are less high. But in fact you see here that in Israel the outlying values are not so high.
12	*D*	I am confused now, I don't understand. Not correct, because if your data…
13	*A*	If everything in Israel is smaller, then you would expect that the outlying values, yes, will be high numbers, since there are few of them; and in the US, the outlying numbers – will be lower, since there are few of the low.
14	*D*	Yes, but this is not correct.
15	*A*	But in fact in the US also - the high are the outliers.
16	*D*	9 and 10.
17	*A*	Right, 10 and 9 are outliers, but 11 is really high.
18	*D*	Correct.
19	*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 ("*everything is smaller*"). 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 the US) 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 the long surnames since the Israeli surnames tended to be short (and vice versa in the US distribution).

A seems to deal with distributions' variability with a plain dichotomous model. In his mental model, he divided 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 US). 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: "*You expect that in Israel the outlying values will be higher than in the US*". 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.

The outlier as a concept provides a mirror to the fragility of their conceptual 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 (e.g., their choice of American outliers).

In sum, 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 clearly present 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 language flow.

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

In the third part of the activity, the students created graphical displays of the data and were asked to use them to compare the distributions. The following dialogue (Video segment 4) took place after they created a series comparison graph (double bar chart) of the two groups (Figure 2).

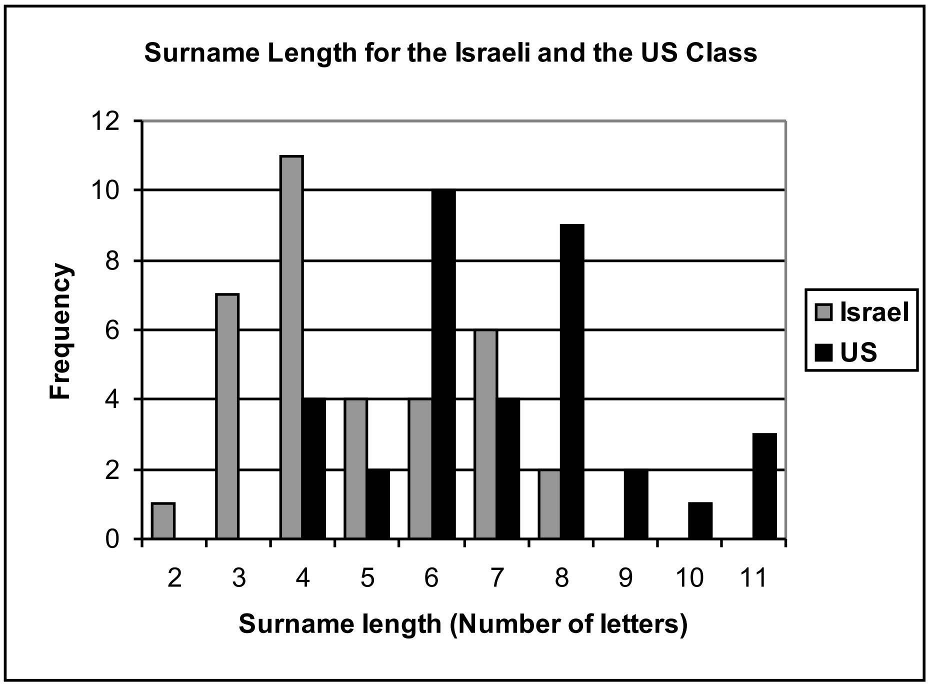

Figure 2. Series comparison graph (double bar chart) of the two groups

1	*D*	[Reading the task] Use the series comparison graph (Figure 2) to describe the emerging trend in the surnames' length of the two countries.
2	*A*	Let's see: The US… usually… no… hold on…
3	*D*	It seems that it's a lower trend in the US.
4	*A*	Not low, it seems about the same in the graph.
5	*D*	Aha… No, higher trend.
6	*A*	Hold on, the US…
7	*D*	Since you do not compare this to that, but rather this to that.
8	*A*	[Cynically] Really!
9	*D*	All right. [Unclear] … seven.
10	*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.
11	*D*	And here.
12	*A*	And here.

13	*D*	They balance each other.
14	*A*	Look, the advantages [height differences] are bigger in Israel. No, not always. Let's ask someone what it means.
15	*D*	I know what it means.
16	*A*	What?
17	*D*	It means that the emerging trend is…
18	*A*	But it is not equal. Look, we said that the US is longer… The US leads in 8, 9, 10, and 11, while Israel leads only in 2, 3, 4, 5, and…
19	*D*	We said that the US names are longer, what's the big deal?
20	*A*	That's right. So, the US leads in the longer names. That's also not a big deal since 2 was not considered at all in the US, while 11 was not considered at all in Israel.
21	*D*	What's the big deal? They were not considered because there are none.
22	*A*	OK, but…
23	*D*	They did not ignore data.
24	*D*	It appears that in Israel the lengths of the lower names are…
25	*A*	No…
26	*D*	The length of the names
27	*A*	In Israel… In Israel…
28	*D*	The lengths of the lower names are…
29	*A*	No. In Israel, the lengths of names with fewer letters have a higher frequency, but in the US, the lengths with… [having difficulties to complete the sentence]
30	*D*	I know how to formulate this. Write down.
31	*A*	No. I first want to hear what you have to say.
32	*D*	OK. In Israel, the frequency of the names with low number of letters…
33	*A*	Relatively low.
34	*D*	… is higher than in the US.
35	*A*	Just a second, low – let's say smaller than 5.
36	*D*	Let's assume so. …is higher than…
37	*A*	No. But there is also one exception here.
38	*D*	The frequency is higher than in the US.
39	*A*	But there is also one exception here.
40	*D*	[Angrily shouting] OK, it's in general! It's a general trend! It's not the trend for the exceptional one.
41	*A*	[Surprised by D's reaction] Buu …
42	*D*	OK. On the other hand, in the US, the trend… the frequency of the long surnames is relatively higher than in Israel.

At First, the students provided conflicting interpretations of the graph (Figure 2), "*It seems that it's a lower trend in the US*", and "*It seems about the same in the graph*". These rather unclear statements are the students' initial attempts to find one global description that accounts for the variability between the groups. This attempt can be considered a progress in compared to their previous interpretations of graphs in the SC, which were mostly local. *D* suggested that their disagreement arose from their different ways of reading the graph, possibly ‘horizontal’ reading – comparing values, vs. ‘vertical’ reading – comparing heights of bars (density, frequency). They then began focusing on comparing the heights of adjacent bars, which was *A*’s method for summarizing the variability between the groups. However, this led them to an impasse: the number of “winning” Israeli and American bars was equal (“*They balance each other*”). 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 finding (1999) 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 this conclusion and was particularly concerned about the distinction between short and long names. This issue, which worried him also at the beginning of the activity, 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 US (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 must ignore the one exception, “*It's a general trend! It's not the trend for the exceptional one*.”

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 US, but the frequency of relatively long names is higher in the US than in Israel*.” Thus their final description of the variability between the groups is based on comparing the frequencies of two subgroups ignoring the deviation from the trend in the center.

The students faced several difficulties in trying to respond to the graphical comparison task. First, they struggled with the issue of where and how the variability within and between the groups is encrypted in the graph (Figure 2). In reading and interpreting the graph they seem to move between noticing ‘horizontal’ variability with a focus on edges (e.g., “*2 was not considered at all in the US, while 11 was not considered at all in Israel*”), and ‘vertical’ variability – focusing on frequency differences between adjacent pairs of bars (“*The US leads in 8, 9, 10, and 11, while Israel leads only in 2, 3, 4, 5, and…*”). Second, they struggled with finding a description that would capture the disorganized nature of these data in a single deterministic formula. Their solution was to divide the distributions to two groups and compare each one of them separately.

Discussion

The above ‘story’ focuses on the first steps of two students learning an EDA curriculum. It concentrates on the way they started to develop views (and tools to support them) that are somewhat consistent with those of EDA experts in the context of comparing real groups of data in a technological environment. This account is offered as a contribution to understanding the process of developing intuitive ways to reason about variability within and between distributions through working on comparing groups. I propose that the analysis of data illustrates the following aspects of student’s emerging understanding and learning of variability in comparing groups.

Development of reasoning about variability

A and *D* started to learn to make sense of general questions normally asked in EDA. Their learning included trying irrelevant answers, 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.

In the raw data stage, the students initially 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 global features of the distributions. Similarly, in the first activity of the SC, *A* and *D* were attentive to the prominence of “local deviations” in data, which also kept them from dealing more freely with global views of data. Only after the teacher’s initial intervention they started focusing on relevant information taking into account the variability in the data. Their reasoning about variability evolved then from observing of differences between two values, followed by generally distinguishing between long and short names, and noticing and informally describing the variability between the groups. They finally arrived at a hypothesis formulation that took into account the variability in the data (“usually, not always”).

In the frequency table task, *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 to their difficulties and incomplete view can be 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.

The students’ insignificant and monotonous use of statistical measures to compare the groups (“*Everything is smaller*”) resembles students’ reluctance to meaningfully use averages 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 and Friel, 1998; Watson & Moritz, 1999). Konold and colleagues (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.

Through their dealing with comparing the outliers between the groups, 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 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 language flow.

They struggled with interpreting the graph (double bar chart, Figure 2) they used to compare the groups. 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 (US 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, in press). 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).

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

- Students' repeated experiments in using different informal methods, mostly local, that capture the variability in the data and between the groups (example: comparing heights of adjacent bars).
- Previous experiences with these data and other sets of data (examples: the comparing strategy based on local differences between adjacent bars, which helped them interpret a time plot in the previous activity; and the dichotomous interpretation of the graph based on the description of the distributions when they handled the table of measures).
- Interactions with the teacher that helped them adopt the statistical perspective but did not instruct them in exactly what to do. The role of the teacher included reinforcing the legitimacy of an interpretation as the right 'kind' in spite of not being fully correct, or simply refocusing attention on the question. These initial steps in an unknown field may be regarded as an aspect of the *enculturation* process (e.g., Schoenfeld, 1992; Resnick, 1988): entering and picking up the points of view of a community or culture. In this process, the teacher is considered as an ‘enculturator’.
- Synthesis of statistical and contextual knowledge: Integration of statistical knowledge and contextual knowledge is considered, by statisticians, “an identifiable fundamental element of statistical thinking” (Pfannkuch and Wild, in press). The context sphere of this problem supported A and D’s statistical reasoning by providing reasonable explanations to the observed patterns in the variation so the students were able to negotiate a

purposeful reason to consider variation in the context of what constitutes a longer surname in the two classes. For example: using the surname 'Raz' to illustrate why American surnames are usually longer.

Instructional implications

The learning processes described in this paper took place in a carefully designed environment that included:

0- 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),
- 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, 1995).

In learning environments of this kind, students meet and work with, from the very beginning, ideas and dispositions related to the culture of EDA (making hypotheses, summarizing data, recognizing trends and variability, identifying interesting phenomena, and handling data representations). Skills, procedures and strategies (e.g., reading graphs and tables, calculating statistical measures, comparing distributions) 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, I propose that what *A* and *D* experienced is an integral and inevitable component of their meaningful learning process (with lasting effects that are now being analyzed on the basis of further data). If students were to work in environments such as the above, the learning would involve the following:

(i) their prior knowledge will (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,
(ii) many questions will either make little sense to them, or, alternatively, will be re-interpreted and answered in different ways than intended, and
(iii) their work will inevitably be based on partial understandings, which will grow and evolve.

This study confirmed that even if students do not make more than partial sense of that with which they engage, appropriate teacher guidance, in-class discussions, peer work and interactions, and more importantly, ongoing cycles of experiences with realistic problem situations, will slowly support the building of meanings.

Several specific implications are to be considered as a result of this study regarding the teaching and learning of variability in the context of comparing groups:

Data and Data Analysis

It is recommended that students be provided with many opportunities to handle data and to investigate data (in the EDA spirit) in order to develop ‘data sense’ including ‘variability-sense’. The focus should be on teaching concepts rather than conventions (in the spirit of Konold, in press). Comparing group tasks should include equal-size and different-size ‘faked’ (educationally engineered), realistic, as well as real data sets. In this study we used real data, which provided opportunities for developing student’s reasoning about variability and comparing groups. However, educationally engineered data could also ‘do the job’ well, providing diverse comparing situations (e.g., comparing two groups in which the mean and median are in opposing directions). It also recommended that a variety of structures of raw data (tables, list, cards, and materials) that can support the recognition and tolerance of variability in various ways be used.

Data Representations

It is recommended that different types of data representations are presented in EDA curricula for comparing groups: conventional graphs (e.g., double bar charts, stem-and-leaf plots, histograms, and box plots) as well as students’ invented inscriptions. New graphical tools are easily available through software and Internet (e.g., Fathom, Tinkerplots, and Mini-Tools). Exposure to the many ways variability is represented in different types of visualizations and simulations will provide students with opportunities to notice, measure, and model variability in data.

Statistical Measures in Comparing Groups

Students should be given meaningful opportunities to consider the different roles statistical measures (center and spread) can play in representing data and in comparing distributions in relation to the idea of variability. Comparing group problem situations can be fruitful arenas for providing the intellectual motivation for developing reasoning about variability and the role of measurers in the comparison. It is recommended to begin with handling raw data before and during the learning of statistical measures. One option to start is by offering students with opportunities to invent measures (and data representations). In comparing groups it is important to provide exemplary data with opposite values of statistical measures. A relatively complex comparing situation can push students to think about the meaning of what they calculate, and carefully check the relevance of their interpretations. It is also recommended to let different models of variability emerge, be compared and discussed in the class, and provide opportunities to develop the reasoning about variability in different contexts and data sets.

Assessment

Multiple challenges exist in the assessment of outcomes of students’ work in such a complex learning environment: the existence of multiple goals for students, the mishmash between the contextual (real-world) and the statistical, the role of the computer-assisted environment, and the group vs. the individual work (Gal & Garfield, 1997). The assessment employed in the SC consists in the use of an extended performance task in similar settings to those exercised during the learning, i.e., open, semi-structured questions, work in pairs, and use of computers. Although shown beneficial in many respects, this method still needs further investigation, in particular to find efficient ways to evaluate the knowledge and dispositions of the individual within a group (Hershkowitz, 1999).

Implication for research

Many research questions remain unsolved in light of this study. Here are just a few examples:

1. Which data representations best support students' development of reasoning about variability in a comparing groups situations?
2. Where do intuitive ideas of variability originate?
3. How does reasoning about variability in data and between distributions develop?
4. What are the roles of particular instructional activities, technological tools, or assessment tasks in developing reasoning about variability?
5. How does one assess reasoning about variability in rich learning environment?
6. What is the role and contribution of quantitative research methods used in statistics education in studying reasoning about variability?

References

Bakker, A., & Gravemeijer, K. (In press). Learning to Reason about Distribution. In D. Ben-Zvi, & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking*. Kluwer.

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* (on CD-ROM). Voorburg, The Netherlands: International Statistical Institute.

Ben-Zvi, D. (In press). Reasoning about Data Analysis. In D. Ben-Zvi, & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking*. Kluwer.

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). 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). 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.

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*. UK: The Teaching Statistics Trust.

Konold, C (In press). Teaching concepts rather than conventions, *New England Journal of Mathematics*.

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*, 193-215. Reston, VA: National Council of Teachers of Mathematics. [http://www.umass.edu/srri/serg/papers/KonoldHiggins72001.pdf]

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,* 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). 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. (in press). Secondary teachers' statistical reasoning in comparing two groups. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking*. Kluwer.

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* (CD-ROM). The Netherlands: International Association for Statistical Education (IASE).

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., 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, 69-100. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Pfannkuch, M., & Wild, C. (in press). Statistical literacy, reasoning, and thinking: goals, definitions, and challenges. In D. Ben-Zvi, & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking*. Kluwer.

Reading, C., & Shaughnessy, M. (in press). Reasoning about variation. In D. Ben-Zvi, & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking*. Kluwer.

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* (CD-ROM), South Africa. The Netherlands: IASE.

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.

Voigt, J. (1995). Thematic patterns of interaction and sociomathematical norms. In P. Cobb & H. Bauersfeld (Eds.), *Emergence of mathematical meaning: Interaction in classroom cultures*, 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* (CD-ROM), South Africa. The Netherlands: IASE.

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

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.

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.

Appendix: Video segments

Segment	Content	Length (min:sec)
1	Formulating a hypothesis	1:56
2	Comparing groups using a frequency table	2:54
3	Range and outliers in comparing groups	2:04
4	Comparing groups using a series comparison graph	2:47
	Total length	**9:41**