"VARIATION-TALK": ARTICULATING MEANING IN STATISTICS

KATIE MAKAR
University of Queensland, Australia
k.makar@uq.edu.au

JERE CONFREY
Washington University in St. Louis, U.S.A.
jconfrey@wustl.edu

SUMMARY

Little is known about the way that teachers articulate notions of variation in their own words. The study reported here was conducted with 17 prospective secondary math and science teachers enrolled in a preservice teacher education course which engaged them in statistical inquiry of testing data. This qualitative study examines how these preservice teachers articulated notions of variation as they compared two distributions. Although the teachers made use of standard statistical language, they also expressed rich views of variation through nonstandard terminology. This paper details the statistical language used by the prospective teachers, categorizing both standard and nonstandard expressions. Their nonstandard language revealed strong relationships between expressions of variation and expressions of distribution. Implications and the benefits of nonstandard language in statistics are outlined.

Keywords: Statistical reasoning; Statistics education; Reasoning about variation and distribution; Mathematics education; Teacher education; Nonstandard language

Do not let calculations override common sense.
Moore & McCabe (1993, p. 121)

1. INTRODUCTION

Recently, researchers in statistics education have been calling for greater emphasis in schools on developing students' conceptions of variation (Moore, 1990; Shaughnessy, Watson, Moritz, & Reading, 1999; Pfannkuch & Begg, 2004). In addition, they argue that too much instruction in statistics has focused on performing statistical operations rather than developing students' thinking about what makes *sense*. One approach to sense-making is through encouraging learners to express their ideas in their own words (Noss & Hoyles, 1996). Russell and Mokros (1990) documented teachers' statistical thinking by attending to their ways of describing data using nonstandard statistical language. Other research has also revealed that learners often articulate concepts of variation using nonstandard language (e.g., Bakker, 2004; Reading, 2004); however these studies have not systematically looked at ways that learners express notions of variation. Even less is known about how teachers express concepts of variation.

In other qualitative studies, we examined preservice secondary mathematics and science teachers' experiences making sense of data as they conducted technology-based investigations (Confrey, Makar, & Kazak, 2004; Makar, 2004; Makar & Confrey, submitted). This paper reports on an exploratory study of preservice teachers' use of standard and nonstandard statistical language when discussing notions of variation. Below we first review the background for the present study, including

Statistics Education Research Journal, 4(1), 27-54, http://www.stat.auckland.ac.nz/serj

the conceptualization of variation, prior research on teachers' and students' knowledge of variation, and research needs regarding meaning construction and language use in statistics education.

2. REASONING ABOUT VARIATION

Although variation is a central component of understanding statistics, little is known about how children (and even less about how teachers) reason with, conceptualize, learn about, and *express* notions of variation. This section begins with a conceptual analysis of variation, then looks at recent research on developing teachers' conceptions of variation, and finally turns to recent work on middle school students' reasoning in this area.

2.1. VARIATION AS A CONCEPTUAL ENTITY

Most uses of the term "variation" in research studies are taken to have a self-evident, common sense meaning, and leave it undefined. Variation is closely linked to the concepts of variable and uncertainty. It is often regarded as a measurement of the amount that data deviate from a measure of center, such as with the interquartile range or standard deviation. Variation encompasses more than a measure, although measuring variation is an important component in data analysis. In considering variation, one must consider not just *what* it is (its definition or formula), or *how* to use it as a tool (related procedures), but also *why* it is useful within a context (purpose).

In simple terms, variation is the quality of an entity (a variable) to vary, including variation due to uncertainty. "Uncertainty and variability are closely related: because there is variability, we live in uncertainty, and because not everything is determined or certain, there is variability" (Bakker, 2004, p. 14). While no one expects all five year-old children to be the same height, there is often difficulty understanding the extent to which children's height vary, and that the variability of their heights is a mixture of explained factors (e.g., the parents' heights, nutrition) and chance or unexplained factors.

In this paper, we will not distinguish between reasoning about variability and reasoning about variation, although Reading and Shaughnessy's (2004, p. 202) definition of variation distinguishes it from variability:

> The term *variability* will be taken to mean the [varying] characteristic of the entity that is observable, and the term *variation* to mean the describing or measuring of that characteristic. Consequently, the following discourse, relating to "reasoning about variation," will deal with the cognitive processes involved in describing the observed phenomenon in situations that exhibit variability, or the propensity for change.

We would argue that developing the concept of variation, beyond just an acknowledgement of its existence, requires some understanding of distribution. In examining a variable in a data set, one is often interested in uncovering patterns in the variability of the data (Bakker, 2004), for example by representing the ordered values of the variable in a graph. The distribution, then, becomes a visual representation of the data's variation and understanding learners' concept of variation becomes closely linked to understanding their concept of distribution. A key goal in developing students' reasoning about distributions is assisting them in seeing a distribution as an *aggregate* with its own characteristics (such as its shape or its mean) rather than thinking of a distribution as a collection of individual points (Hancock, Kaput, & Goldsmith, 1992; Konold & Higgins, 2002; Marshall, Makar, & Kazak, 2002; Bakker, 2004).

In thinking about what the notion of distribution entails, one might think of the associated measures, properties, or characteristics of a distribution—for example, its mean, shape, outliers, or standard deviation. These entities in isolation, however, do not capture the desired concept and can aggravate the focus on individual points. In addition, they lack the idea that we want to capture a distribution *of* something. Pfannkuch, Budgett, Parsonage, and Horring (2004) cautioned that students often focus on characteristics of a distribution, but forget to focus on the meaning of the distribution within the context of the problem. In comparing distributions, for example, they suggested that students should first look at the distributions and compare centers, spread, and anything noteworthy, but then should be asked to draw an evidence-based conclusion, using probabilistic rather than deterministic language, based on their observations.

2.2. TEACHERS' REASONING ABOUT VARIATION

Teachers' mind-set and conceptions about data have been of great concern in the research community for over a decade (Hawkins, 1990; Shaughnessy, 1992). Despite this, Shaughnessy (2001) reports, "[I am] not aware of any research studies that have dealt specifically with teachers' conceptions of variability, although in our work teaching statistics courses for middle school and secondary school mathematics teachers we have evidence that many teachers have a knowledge gap about the role and importance of variability."

Three very recent research projects have focused specifically on teachers' notions of variation and distribution. Hammerman and Rubin (2004) report on a study of a two-year professional development project that gave secondary teachers an opportunity to analyze data with a new data visualization tool. The software Tinkerplots™ (Konold & Miller, 2004) allows learners to create their own data representations through sorting, ordering, and stacking data. They found that teachers did not choose to use measures of center in comparing distributions even though those measures were readily available in the software, but rather chose to compare groups by comparing slices and subsets of distributions to support compelling arguments with data. In addition, Hammerman and Rubin noted that teachers were dissatisfied with representations such as histograms and box plots that would hide the detail of the underlying distribution. From their study, we can begin to see the potential that new visualization tools have for helping teachers to construct their own meaning of statistical concepts.

In another study, Canada (2004) created a framework to investigate prospective elementary teachers' expectation, display, and interpretation of variation within three statistical contexts: repeated sampling, data distributions, and probability outcomes. His study took place in a preservice mathematics course developed to build the teachers' content knowledge of probability and statistics through activities that emphasized hands-on experiments involving chance and computer-generated data simulations. Canada's study found that initially, although the prospective teachers had a good sense of center, they found it difficult to predict realistic spreads in distributions from collected data or produced from random sampling and probability experiments. Their predictions for underlying distributions included expectations of way too little variation, overly extreme variation, or unrealistically symmetrical distributions. In addition, many of the teachers took an initial stance that 'anything goes' in random experiments and felt that they could therefore make no judgment about the distribution of outcomes. After the course, the teachers demonstrated stronger intuitions about variation - their predictions were much more realistic and their expectation of variation more balanced. In addition, the teachers' descriptions of distributions were more rich and robust in their recognition of variation and distribution as an important concept, for example by referencing how the distributions were clustered, spread, concentrated, or distributed. Canada's study hints at a potential link between the elements that teachers focus on when describing distributions and their intuition about variation and uncertainty in data.

In previous research on statistical inquiry with practicing middle school teachers, we found that teachers, like their students, often begin examining data distributions by focusing on individual points (Confrey & Makar, 2002). Yet in a meaningful context (e.g., student test scores), the teachers in our study constructed a need for examining variation in a distribution by initiating a discussion of how the distribution of students' abilities affected their choice of instructional strategies. We noticed, however, that when the teachers did not adequately construct more complex concepts, like sampling distributions, they would tend to use statistical tools mechanically, without carefully examining their relationship to the data (Makar & Confrey, 2004).

In summary, although much concern has been expressed about the over-emphasis on procedures in teaching statistics, the studies described above highlight some important ways to improve what Shaughnessy (1992) calls teachers' lack of intuition about stochastics. The main ideas these studies bring out is that teachers themselves need to learn statistical concepts in an environment much like the one recommended for students – one that is active, technology-rich, involving authentic data, and offering plenty of opportunities to build their conceptions through experiences with data. Particulars about how teachers build these conceptions still needs further research.

2.3. STUDENTS' REASONING ABOUT VARIATION

It is useful to briefly review research on students' development of concepts of variation, as it may provide insight into potential ways to build teachers' conceptions of variation. First, teachers frequently possess similar reasoning to their students (Hammerman & Rubin, 2004; McClain, 2002). Second, understanding students' conceptions of variation can help teachers to plan instruction. Finally, the research in this area has uncovered new insights into students' intuitions about variation without formal procedures and terminology. These insights may provide ideas for new methods of professional development for teachers.

A common thread in the research on teaching concepts of variation and distribution is a recommendation to focus not only on the characteristics of distributions, but on their *purpose*. A well-documented approach to this is through comparing groups (see for example, Cobb, 1999; Watson & Moritz, 1999; Biehler, 1997; Makar & Confrey, 2004; Hammerman & Rubin, 2004). Pfannkuch et al. (2004) found that the descriptions of distributions given by secondary students sometimes included characteristics of the distribution that were disconnected from the context or meaning of the problem. For example, when taking note of the variability of the data, half of the students compared the ranges, which was not relevant to their question, and most of the students focused on comparing measures of center or extremes. Pfannkuch and her colleagues hypothesized that the instruction the students received focused not on drawing meaningful conclusions, but rather on comparing features of the box plot. They recommended that instruction concentrate not on *how* to compare centers, but *why* one should do so. Reading and Shaughnessy (2004) also found that tasks which asked students not just for descriptive information but also for *explanations* offered greater possibilities for insight into students' reasoning about variation.

In developing students' conception of statistics, Konold and Pollatsek (2002) argue that "the central idea should be that of searching for a signal and that the idea of distribution comes into better focus when it is viewed as the 'distribution around' a signal" (p. 262). Bakker and Gravemeijer (2004) on the other hand hold that "reasoning with shapes forms the basis for reasoning about distributions" (p. 149). By developing a lens of seeing the distribution as an entity, one can then look at statistical measures as characteristics of the distribution rather than as calculations from individual points (Bakker, 2004).

Bakker (2004; Bakker & Gravemeijer, 2004) also found that with Cobb's Minitool software and an innovative learning trajectory he was able to encourage his middle school students to think about variation and towards a distribution-view of data. Initial discussions with students focused on the middle "bump" of the distribution, but later the "bump" came to represent the whole distribution. Konold and his colleagues (2002), who studied middle school students, emphasized that in problem solving with data, the middle bump was a frequently identified portion of a mound-shaped distribution, and termed this central bump a *modal clump*. Bakker (2004) also found that rather than just focusing on the central region of a distribution, students also tend to divide distributions into three categories: low, middle, and high. Hammerman and Rubin (2004) saw the teachers in their study take a similar tactic in comparing distributions. These findings may indicate that dividing distributions into three pieces may be a more natural way for students to examine a distribution than by dividing it into four sections, as in the box plot.

The studies reported above suggest that if learners are provided with relevant contexts, concrete experiences, complex tasks, adequate time and support, and appropriate tools to build statistical concepts, then their understanding appears to be much more robust. One notable common feature of these studies is that they developed learners' notions of variation through constructing a purpose for variation, often building on the learners' own language for describing and interpreting what they were seeing.

3. CONCEPT DEVELOPMENT

Below we briefly discuss some aspects of how learners construct meaning of data as well as issues surrounding the use of learners' everyday or nontechnical terminology. These issues affected the design of the present study and the lens we used to interpret prospective teachers' articulation of variation. The need to refer both to standard statistical terminology and to nonstandard or nontechnical statistical terminology arises because one cannot assume that if a respondent can repeat the definition of a concept, she has necessarily assimilated this concept. For example, although a student may be taught the concept and formula for standard variation and can even use this term as part of class discourse, this does not imply that they are "seeing" variation in what is being measured.

3.1. CONSTRUCTING MEANING

Because students are frequently taught definitions and procedures without first developing their own intuition and meaning about the concepts underlying them, premature instruction of formal terminology and rules can inhibit students' own sense-making (Flyvbjerg, 2001; Schoenfeld, 1991; Boaler, 1997). Formal mathematical procedures, terminology, and symbolism are critical for developing advanced levels of mathematical understanding in that they can provide efficient paths to problem-solving, focus attention on particular aspects of a problem, and open new levels of understanding of the concepts represented by the terms or symbols. However, the emphasis must be on building meaning, not simply assuming that standard procedures or terms can themselves carry the meanings of underlying concepts.

Shaughnessy (1992) notes that students' and teachers' lack of intuition about stochastics is a critical barrier to improved teaching and learning in statistics. Fischbein's (1987) description of intuition matches well with the kind of thinking that we believed should be developed in teachers.

> *Intuitions are always the product of personal experience, of the personal involvement of the individual in a certain practical or theoretical activity*. Only when striving to cope actively with certain situations does one need such global, anticipatory, apparently self-consistent representations. The development of … intuitions implies, then, didactical situations in which the student is asked to evaluate, to conjecture, to predict, to devise, and check solutions. In order to develop new, correct probabilistic intuitions, for instance, it is necessary to create situations in which the student has to cope, *practically*, with uncertain events (p. 213, italics in original).

For building intuition, Fischbein further argues that the role of visualization "is so important that very often intuitive knowledge is identified with visual representations" (p. 103). Insight can also be gained into learners' sense-making by focusing on how they use their own words to explain relationships, constructs, and processes (Noss & Hoyles, 1996). One might infer from these last three quotes that intuition in statistics is built through the development of visualization, articulation, and representation of data distributions within personally compelling contexts. Intuition about variation, then, may be fostered through a lens of "seeing variation" (Watkins, Scheaffer, & Cobb, 2003, p. 10), or Moore's (1990) acknowledgement of the omnipresence of variation. These insights may be gained through attending to students' own words for describing what they are seeing. These words are often less technical and contain elements of nonstandard language.

3.2. NONSTANDARD LANGUAGE USE

Noss and Hoyles (1996) state that focusing on mathematical meaning moves us to consider how learners *express* mathematics rather than how they *learn* it. The difference is in the locus of the concept. If we focus on how students learn statistics, it implies that statistical concepts are ontologically fixed and that the goal of learning is to impart *a priori* knowledge from teacher to student. However, by turning that around to foster learners' dynamic conceptions of statistical concepts, we acknowledge that these concepts are not passively received, but rather are actively and socially constructed by the individual.

Developing understanding oftentimes requires the use of nonstandard terminology. Unfortunately, when students do use their own language to make meaning, teachers often do not recognize nonstandard ways of talking (Lemke, 1990). Biehler (1997) attributes some of the difficulty in communicating and understanding relationships in data to the lack of formal language we have in describing distributions beyond the level of statistical summaries:

> The description and interpretation of statistical graphs and other results is also a difficult problem for interviewers and teachers. We must be more careful in developing a language for this purpose and becoming aware of the difficulties inherent in relating different systems of representation. Often, diagrams involve expressing relations of relations between numbers. An adequate verbalization is difficult to achieve and the precise wording of it is often critical. There are profound problems to overcome in interpreting and verbally describing statistical graphs and tables that are related to the limited expressability of complex quantitative relations by means of common language (p. 176).

Biehler's work implies that the current focus on statistical summaries in describing distributions is inadequate and that research needs to be improved in the area of developing statistical language for interpreting more robust relationships in data.

4. DESIGN AND METHODOLOGY

4.1. APPROACH

This exploratory study was designed to gain insight into the ways that prospective secondary mathematics and science teachers express or discuss notions of variation when engaged in a purposeful statistical task. Based on the literature reviewed, it was anticipated that respondents will use standard or conventional descriptions and terms, as well as nonstandard descriptions and informal language, and the goal of the study was to document the different types of language used. Respondents were interviewed twice using an identical task, in the first and last week of a fifteen-week preservice course on assessment, which included an embedded component of exploratory data analysis. The task given during the interview asked teachers to compare two distributions of data relevant to the context of teaching, in terms of the relative improvement in test scores of two groups of students.

Although interviews were conducted at the beginning and end of the course, the purpose of the study was not to evaluate the effectiveness of the course or to compare performance before and after instruction. Some comparisons will be made of language use before and after the course, but given the small number of respondents, these comparisons should be interpreted with caution. The primary interest was in categorizing and describing the language respondents use to describe variation in the data. Any change that may have occurred in respondents' language use simply enriched and extended the range of responses available for analysis.

4.2. SUBJECTS

The respondents for the study were secondary mathematics and science preservice teachers at a large university in the southern United States enrolled in an innovative one-semester undergraduate course on assessment designed and taught by the authors (Confrey et al., 2004). Twenty-two students began the course, but four withdrew from the course before the end of the semester. In addition, some data from one subject was lost due to technical malfunction, leaving seventeen subjects with a complete set of data. The seventeen subjects ranged in age from 19 to 42, with ten students of traditional college age (19-22), five aged 23-29, and two students over 30 years of age. Of these, three were male and fourteen female, nine were mathematics majors and eight were science majors (predominantly biology). About 60% of the class was Anglo, with the remaining students being of Hispanic and African-American ancestry. The students had varying backgrounds in statistics: eight had not previously studied statistics, while five had previously taken a traditional university-based statistics course either in the mathematics department or in a social science department. The remaining four had not taken a formal course, but had previous experience in statistics as a topic in one of their mathematics or science courses.

4.3. SETTING

The study took place at the beginning and end of a one-semester course that integrated ideas of assessment, data analysis, equity, and inquiry—themes identified as critical but missing from preservice education (National Research Council, 1999; 2000; 2001). The purpose of the course was to give the preservice teachers some background in classroom and high-stakes assessment, develop their statistical reasoning, gain experience using technological tools to interpret student assessment results, and to introduce them to issues of equity through examining data. In the final month of the course, the prospective teachers conducted their own data-based inquiry into an issue of equity in assessment.

The prospective teachers were guided through several investigations that built an atmosphere of interpretation of data rather than the development of formal theoretical foundations. The inclusion of the dynamic statistical software Fathom™ (Finzer, 2001) was critical as a learning tool rather than a traditional statistical package aimed at statisticians and statistics students. The statistical content of the course was comprised of an overview of data graphing (histograms, box plots, dot plots), descriptive statistics (mean, median, standard deviation, interquartile range, distribution shapes), linear regression (association, correlation, least-squares, residuals), and a brief introduction to sampling distributions and inference (through building of simulations). The statistical content was developed as a set of tools to gain insight into data rather than as isolated topics.

4.4. INTERVIEW TASK AND PROCEDURE

Respondents were interviewed by the first author during the first and last week of the course, using the same task in both interviews. Most interviews lasted between ten and twenty minutes. The task was set in the context of an urban middle school trying to determine the effectiveness of a semester-long mathematics remediation program (called 'Enrichment') for eighth-grade (13-14 years old) students considered in need of extra help preparing for the state exam given each spring. To decide if the Enrichment program was working, the school compared student scores from their seventh grade state exam score in mathematics to their scores on a practice test given near the end of eighth grade. Respondents were shown a pair of dot plots (Figure 1) of authentic data taken from students in an Enrichment class (upper distribution) and a regular eighth grade class (lower distribution). Respondents were asked initially to compare the relative improvement of students in the two groups, and then were probed if their responses needed clarification. It is important to note that the data in Figure 1 represent the *change* (difference) in scores between the two assessments, i.e. numbers on the x-axis are positive when scores improved, and negative when scores declined. (Data points in red (shaded dots) highlighted those students classified as economically disadvantaged; this was used for another question in the interview related to equity but not pertinent for this study.)

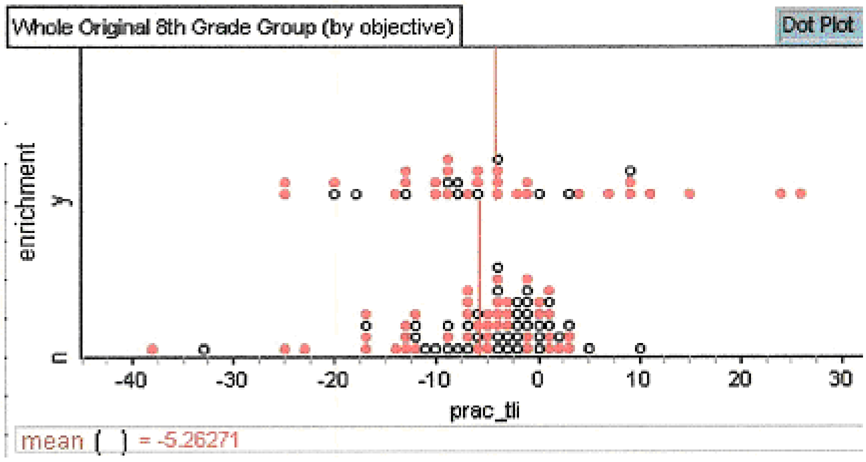

Figure 1. Graph shown to subjects during the interview task.

The preservice teachers were walked through particular elements of the graphs both in the pre-interview (January) and post-interview (May). It was explained that the data on the horizontal axis represented the improvement of each student from the seventh grade state exam to their eighth grade practice test. The mean improvement of each group, marked by a vertical line segment in each distribution, was pointed out. The overall mean (displayed as -5.26271 at the bottom of the graph) was interpreted to respondents as the average improvement of the entire eighth grade (that is, both groups combined) and it was also noted that the overall mean improvement lay in the negative region, meaning that generally the students had performed less well on their eighth grade practice test than they did on the seventh grade state exam.

In our choice of task we considered it important to make the context and task as authentic as possible in order to examine the prospective teachers' responses in a situation close to what they would encounter in their professional life. Therefore, rather than use hypothetical data constructed to emphasize a particular aspect of the distributions, actual data from a local school was used. These somewhat "messy" data made the task, and hence our analysis, more difficult. Yet, because authentic school data is rarely "clean" this setup provided the benefit of examining how the prospective teachers would interpret actual (and messy) school data. We recognize that unintended elements of this particular representation may have influenced the subjects' thinking about the task (Kosslyn, 1994). We had the teachers consider the "improvement" of students since improvement is a natural construct in teaching, and means were marked because this is a common method for comparing distributions, giving the subjects a potential starting point for their discussion. In addition, it allowed for insight into whether the prospective teachers would interpret a small difference in means deterministically, or if they would expect some variability between the means (Makar & Confrey, 2004).

4.5. ANALYSIS

The videotaped interviews from all twenty-two subjects were transcribed and then the content analyzed and coded to find the categories of concepts that emerged from the data. General categories were initially sought through open coding to isolate concepts that might highlight thinking about variation and distribution, and those passages identified by these codes underwent finer coding resulting in eighteen preliminary categories. Since codes were not predetermined, but rather allowed to emerge from the data, this portion of the analysis was not linear and underwent several iterations of coding, requiring a back-and-forth analysis as codes were added, deleted or combined. Commonalities and differences were examined in passages coded under each node to better describe and isolate the category, determine dimensions and distinctions among participants' descriptions, and locate exemplars. Although the data from all twenty-two original subjects were coded to determine categories, dimensions, and exemplars, only the seventeen subjects with complete data sets (both pre- and post-interviews) were used in quantitative descriptions.

5. RESULTS

In this section we first overview key categories of standard terms and concepts that the respondents, prospective teachers, used when comparing distributions. Next, categories of nonstandard language and terms are described, referring to two separate but overlapping areas: variation (e.g., *spread*) and distribution (e.g., *low-middle-high*, *modal clumps*). We later refer to these two types of non-standard categories as "variation-talk". Finally, relationships found within and between the categories of standard and nonstandard language categories will be summarized. As explained, the primary goal of the study is to provide a rich description of prospective teachers' language when discussing variation; hence, the information from the January and May interviews is usually combined. Changes in respondents' articulations from the first interview to the interview conducted after the course are noted as well but should be interpreted with caution in light of the small number of respondents and other factors noted later.

5.1. STANDARD STATISTICAL LANGUAGE

This subsection describes the conventional statistical language used by the respondents. Table 1 summarizes the percentage of respondents articulating each category of standard statistical terms and descriptions in their responses. As can be seen in the table, multiple types of standard expressions were used by respondents (i.e., percentages sum to more than 100%) and overall, nearly every subject included at least one type of standard statistical description in their response (94% in January and all respondents in May). Most respondents included the proportion (or number) improved or the mean in their descriptions and the inclusion of standard statistical terms in their responses increased in all categories by the end of the course.

Table 1. Percentage of respondents using standard statistical language (N=17)

Category	January	May
Proportion or number improved	59%	65%
Mean	53%	88%
Maximum/Minimum	29%	41%
Sample size	18%	47%
Outliers	18%	41%
Range	12%	47%
Shape (e.g., skewed, bell-shaped)	12%	35%
Standard deviation	0%	12%
Overall	94%	100%

Proportion or number improved

The most common comparison the respondents made in January was through reporting on the students in each group whose scores improved or dropped, with three respondents describing improvement as the sole element in their comparison. The prevalence of discussing improvement is not surprising given that the data in the task measured the improvement of students' scores.

In most cases, respondents split the groups into two—improved or not improved—as exemplified by Hope (all names are pseudonyms):

Hope: Some people have improved. A lot of people. (Jan)

In May, mention of the proportion of students improving persisted in the respondents' comparisons of distributions. The respondents were more likely to quantify their descriptions and none of them relied on proportions as their sole piece of statistical evidence.

Charmagne: There seem to be, like a split between, um, those who improved and those dropped, like, sort of, off the 50-50 split. (May)

One could argue that a focus on the proportion of students who improved does not necessarily imply that respondents are visualizing the variation in the data, nor seeing the distribution as an aggregate.

Means

Only about half of the respondents interviewed used the average in their descriptions in January, despite the fact that the means for each group were marked on the figure and also pointed out by the interviewer when describing the task. Use of the mean in comparing the distributions ranged from a brief mention to a major focus of their discussion. For example, two of those who mentioned the mean did not use any other statistical descriptions to compare the distributions:

Mark: Well, it looks to me like, uh, the group that did the Enrichment program overall has a better, uh, improvement even though it's not really even- [an improvement].

I: Okay. … And what are you basing that on?

Mark: Uh, cause you. I think you said that this line was the mean? … So, uh, I was looking at that. (Jan)

José: It seems about even. I mean, they didn't decrease by that much, compared to the other [group]. … I don't even know what that would be, a point between their mean and their mean? (Jan)

In previous work (Makar & Confrey, 2004), the authors noticed that while some respondents had a deterministic view of measures, others indicated some tolerance for variability in means, as did two respondents in this study who recognized the effect a small sample could have on the variability of the mean. The first excerpt below comes from Angela, a teacher with no formal training in statistics, whereas the second teacher, Janet, was a post-graduate student with a strong background in statistics.

Angela: Um, well, it's, I guess, obvious, I guess that. As this group [Enrichment], they did improve more, just I mean, because their average is better. But it's not a huge dramatic difference. … I mean, there's not as many in the Enrichment program [as the non-Enrichment] and they did improve more, but yet, I mean, I mean out of a smaller group of number. So their mean, I mean, comes from a smaller group. … I mean, if there were more kids, their average might have been different. (Jan)

Janet: So the Enrichment class did have a higher mean improvement, higher average improvement, uh, but they had a smaller class. Um, I don't know what else you want me to tell you about it.

I: You said they had a smaller class, is that going to have any-

Janet: A smaller sample size can throw things off.

I: How's that?

Janet: (laughs) Um. The, with a, a larger population the outliers have less of an effect on the, on the means than in a smaller sample. So it doesn't, um, I don't remember how to say it, it doesn't, uh, even things out as much. (Jan)

Janet's initial statement "I don't know what else you want me to tell you about it" may imply that she saw a difference in the means, but little else worth discussing. Kathleen, who had also used some statistics before in science, recalled comparing means there:

Kathleen: The mean [Enrichment] was a little bit higher than the, the group who didn't, who didn't take the Enrichment class. And I don't know if that would be statistically higher, but-

I: What do you mean, 'statistically higher'?

Kathleen: Like if you, if you ran statistics on it. Like a t-test or something.

When pressed further, Kathleen went on to explain in more detail:

Kathleen: If you, um, if you normalize the data, and um, brought them in together. In fact, once you normalize it for the number of students in this case [Enrichment group] versus the number of students in this case [non-Enrichment group] and brought them, like, closer together for the, for the number of students, and normalized it, then I think the difference [in the means] wouldn't be as great. (Jan)

Although through further probing she was unable to articulate what she meant by "normalize the data", it seems likely that Kathleen was referring to the dependence of sample size on key outcomes of the Central Limit Theorem to compare means with sampling distributions. In the May interviews at the end of the course, nearly all of the respondents mentioned the means, often with more specificity:

Anne: Well, it looks like the students in the Enrichment class, on average, um, improved, or didn't decline as much as the ones in the regular class. Um.

I: And what are you basing that on?

Anne: The means. Uh, the regular class is down by negative, uh, seven, six, minus six. And the Enrichment on average is at minus, um, is that three? (May)

While mean and percentage improvement are important considerations when determining the effectiveness of a class targeted to help students improve their test scores, our hope was that the respondents would do more than just reduce the data and compare means or percentage improvement as their sole method in determining how well the Enrichment program may have worked. Instead, we sought a more robust understanding of the context and an examination of the whole distribution in describing their comparisons.

Outliers and Extreme Values

Another common notion in comparing the relative improvement of each group described by the respondents at the beginning of the course arose through examination of outliers and extreme values. For example,

Andre: Well, it seems like with a few outliers here and a few outliers here, they're pretty similar, um, in terms of how much they changed. (Jan)

Andre had previously studied statistics. The descriptions by other students in the course with no statistics background were less precise:

Gabriela: There's only, like these two out here that have actually, like, greatly intensely improved. (Jan)

Gabriela focused on not just the criterion of *whether* students improved, but qualified it with *by how much*, suggesting that she was seeing the upper values of the distribution and not just whether or not students improved. Note that Andre and Gabriela are not focusing on individual points, but on a set of values at the high or low portion of the graph (e.g., those who "greatly intensely improved").

Shape

The interview task likely did not illicit a need to formally describe the shape of the distributions (e.g., skewed, normal), so their summary here is brief. Traditional shape descriptions were unlikely in January and only somewhat more common in May:

Christine: The non-Enrichment group seems to be skewed to the left. Uh, which means that any outliers that they do have are in the way negative region. Um. The Enrichment group seems to be more normal. It's slightly skewed to the right, but not quite. (May)

Standard Deviation

No one in January and only two of the subjects in May made any mention of standard deviation, a traditional measure of variation, despite the fact that it was discussed in class and included in a homework assignment. In one of these cases, a teacher mentioned standard deviation, but not for any particular argument except to state its relative size in each distribution:

José: Probably the standard deviation is going to be, like, really large on this [Enrichment], compared to that [non-Enrichment], because this is pretty spread out pretty far. (May)

Another prospective teacher indicated that she knew the term, but implied that she did not see it as useful in comparing the relative improvement of the students in the Enrichment program with those who were not, stating a few minutes into her interview:

Charmagne: Um. Yeah. There is more variation in the Enrichment class. This seems to be kind of mound-shaped also. So. I mean. Probably like 65% is in one standard deviation, [laughs] I'm just babbling now. Did I answer the question yet? (May)

From these, the only two examples of the subjects mentioning standard deviation, it would appear that the notion of standard deviation as a measure of variation did not hold much meaning for the respondents. Both of these excerpts pair the use of standard deviation with other less standard expressions that described the variation ("pretty spread out") or shape ("more variation" vs. "kind of mound-shaped"). This may imply that these less conventional descriptions of variation aided them in making meaning of standard deviation.

Range

It was often difficult to tease out notions of variability from descriptions of measures of variation, particularly when the respondents used terms like "range". Ten respondents used the word "range" during the interviews (2 in January and 8 in May, with no respondents in common). In almost every case, their use of it was linked to notions of either measure or location (an interval).

Carmen: The Enrichment class definitely had a better performance since most people are concentrated in this area, whereas you have a wider range and even a very good amount of points that they improved on. … I think it is working, yeah. Because you have a just wider range, whereas everyone was kind of close in on their improvement here. Uh. With the wider range, um, I would say it's working at least for some of the students. Because in the non-Enrichment, no one seemed to improve that much. … Because there's not a range here [the upper portion of the scale]. (Jan)

Carmen's use of the word "range" four times in this passage communicates range as an interval of values in the distribution rather than as a measure. First, she contrasted "wider range" with "concentrated in this area" and "close in … here" giving the impression that she was expressing that the data were spanning a greater part of the scale at a particular location. Next, her suggestion that the wider range implied it was "working at least for some of the students" indicates that it was located in the upper part of the scale, unlike the data for the non-Enrichment group where she said "there's not a range here [the upper portion of the scale]." Brian also used the term "range" to indicate the scale:

Brian: It seems pretty evenly distributed across the whole scoring range. (Jan)

In the May interviews, the use of "range" was more common than in January, even though it wasn't a term we made use of formally in the class. In almost every case, the term "range" either meant an interval, as in the case of Carmen and Brian above, or a measure, like April:

April: The distribution, um, like the lowest the scores in the distri-, the length of the distribution, see this one starts, it's. [pause] … This one is about negative, almost negative forty, I'd say. And this one goes up to ten. So, that's about 50. And this one's about negative 25 and this one's right about 25, a little more, so that's about 50. So, I guess the range is about the same. (May)

Gabriela: There's a lot less of them improving in the Enrichment program, but it's still better that they go off by about five or ten points … then for them to have gone off by forty or twenty. Still kind of in this range. (May)

April's use of the term range is more numeric whereas Gabriela's use appears to indicate a segment of the scale. One difficulty may be that in school mathematics, the term "range" is usually related to a function and defined as a set, almost always an interval on the real number line. In statistics, however, the term "range" is a measure—the absolute difference between the minimum value of a distribution and the maximum value. By using the same term to indicate a set and a measure, we begin to see where the distinction between objects and measures become murky in statistics. In school, the distinction between a geometric object (like a polygon), a measure of it (its area or perimeter), and a non-numerical attribute or categorization of it (closed or convex) is made clear. In teaching statistics, we have not emphasized a clear distinction between an object (e.g. a distribution), a measure of it (its mean or interquartile range), and an attribute (e.g. its shape). This may cause some problems for students trying to make sense of statistical concepts.

5.2. NONSTANDARD STATISTICAL LANGUAGE

This subsection documents the phrases used by the respondents to articulate statistical concepts which could not be categorized as standard statistical terminology. Two dimensions of nonstandard statistical language emerged from the observations made by the respondents (Table 2): *spread* and *distribution chunks*. Nonstandard statistical phrases were categorized into one of these two categories (i.e., they were mutually exclusive), however the dimensions of spread and distribution, as we will show in subsection 5.3, were related. Similar to the results of the use of standard statistical language in the previous subsection, many respondents included both types of nonstandard statistical descriptions (i.e. percentages sum to more than 100%). As can be seen in Table 2, most respondents (53%) included some mention of distribution chunks in their interviews in January and this percentage increased somewhat in May. Although few respondents (35%) discussed the spread of the data in the January interviews, this percentage increased markedly in May. Overall, in both the January and May interviews, the majority of respondents included some kind of nonstandard statistical language in their responses and this percentage increased from the beginning to the end of the course.

Table 2. Percentage of subjects making observations in two dimensions of nonstandard language (N=17)

Dimensions	January	May
Spread	35%	59%
Distribution chunks	53%	65%
Overall	59%	76%

***Expressions of variation:* Clustered *and* spread out**

Here we will document the nonstandard statistical language used by respondents in the interviews that capture their articulation of variation. Their words encompassed a diverse range of language, but the concepts they articulated were fairly similar.

Some respondents used the word *clustered* to describe the relative improvement of each group, like Andre and Margaret, both older college students with previous statistical experience:

Andre: I don't know what to make of this, actually, because as far as, like, it seems to me to support little difference between the Enrichment group and the other group. Because. Um. Both groups are kind of *clustered* around the same area. (Jan)

Margaret: [The non-Enrichment data] are more *clustered*. So where there's little improvement, at least it's consistent. This [Enrichment] doesn't feel consistent. First impression.

I: And you're basing this on?

Margaret: The *clustering* versus, it's like some students reacted really well to this, and some didn't. But it's more *spread out* than this grouping. (Jan)

Andre's use of the term "clustered" highlights his observation that the location of the modal clump in the two distributions overlapped. On the other hand, Margaret's initial description of "clustered" is paired with a notion of consistency, a concept closely related to variability (Cobb, 1999). She goes on to include it in contrast to being "more spread out", a phrase commonly associated with variation (Canada, 2004; Bakker, 2004; Reading, 2004). Five other respondents made use of the phrase "spread out" during the interviews. A few excerpts are given below.

Brian: It seems to be pretty *evenly distributed* across the whole scoring range. Like from about 30 to [negative] 25, it appears pretty *evenly spread out*. (Jan)

Janet: They seem pretty *evenly distributed*, … fairly *evenly spread out*. (Jan)

April: This distribution is more *skewed to the left* and this one is more *evenly spread out* … more of an *even distribution*. (May)

The respondents who used the phrase “spread out” also accompanied it with the word *evenly*. They may be expressing that the data were dispersed fairly equally throughout the scale of the distribution, particularly given its common pairing with the phrase *evenly distributed* or *even distribution* in all three cases above. This context gives *spread out* a meaning related to the *shape* of the distribution, particularly given the contrast April made between *skewed left* and *evenly spread out*. Carmen’s description below gives the phrase a similar meaning:

Carmen: It’s more *spread out*, the distribution in the Enrichment program, and they’re really kind of *clumped*, um, in the non-Enrichment program. (May)

Given this interpretation, we turn back to and expand Margaret’s excerpt to re-examine her use of the phrase *spread out* contrasted with *clustered*, which is similar to Carmen’s *clumped* above:

Margaret: It’s interesting that this is not, that this is not, um, this is much more *spread out* than this group, so. I mean, first impressions. … These are more *clustered*. So where there’s little to no improvement, at least it’s consistent. This doesn’t feel consistent. First impression.

I: And you’re basing that on?

Margaret: The *clustering* versus, it’s like some students reacted really well to this, and some didn’t. But it’s more *spread out* than this grouping. Is, am I saying that okay? (Jan)

The term *spread out* in all these cases appeared to point to an attribute of the distribution akin to shape; we would argue that *clustered* and *clumped*, phrases that accompanied *spread out*, could also be attributes that describe the shape of the distribution, similar to Konold and his colleagues’ (2002) notion of *modal clump*.

How is *the spread* related to *spread out*? Given their similarity, how does its use as a noun compare with its use as an adjective?

Carmen: If you were just to, you know, break the distribution in half. Kind of based on the, the scale, or *the spread* of it, it just seems that, you know, the same amount of students did not improve in both. (May)

Here, Carmen indicates that she is using the noun *the spread* as an indication of length by her pairing of *the spread* with *the scale*. Rachel, in her interview in May, first compared the means of the Enrichment and non-Enrichment groups, then turned to the range, and finally, below, finishes the interview by discussing the way the distribution looked:

Rachel: It’s more *clumped*, down there in the non-Enrichment. And kind of more *evenly distributed*. [Points to Enrichment] … Let’s see. Range and *spread*. That’s what I always first look at. And then average. (May)

Margaret: [The Enrichment distribution] has a much wider *spread* or *distribution* than this group [non-Enrichment]. (May)

In all three of these cases—the only ones where *spread* is used as a noun—Carmen, Rachel, and Margaret convey a meaning of spread as a visual (rather than numerical) attribute of a distribution. Rachel uses the word *spread* to categorize her description of two contrasting terms *more clumped* and *more evenly distributed*. Note that she also distinguishes her notion of *spread* as different than *range* (as a measure), which she discussed earlier in the interview. Margaret directly pairs the word *spread* with its apparent synonym *distribution*, although she likely was using the word distribution in a more colloquial sense.

Another phrase that conveyed similar meaning to *spread* was *scattered*, as used by three respondents:

I:	The first thing I want you to do is just to look at those two and compare the two in terms of their relative improvement or non-improvement. We're trying to determine if the program is working.
Hope:	Well, it's doing something.
I:	What do you mean?
Hope:	I mean, they're more *scattered* across, these guys. … It's helping a little.
I:	Okay. And you're basing that on?
Hope:	On. Well, there's more *grouped* right here. … But you have guys spanning all the way out to here, so it's helping. … It's helping, it's *scattering* them more, it seems. Instead of them all having, so *grouped* together. (Jan)

Hope's descriptions are akin to those we heard above with phrases like *spread out* and *clustered* and *clumped*. Janet's and Anne's uses were similar:

Janet: There's Economically Disadvantaged kids pretty much *scattered* throughout both graphs. (May)

Anne: I mean these are all kind of *scattered* out almost evenly. Whereas these are more *bunched* up together. (May)

If we substitute *scattered* with *spread out* above, notice how the meaning doesn't appear to change. Also note Anne's contrast of *scattered out* with *bunched up*. One more pairing may also be included here:

June: It seems that the people that weren't in the Enrichment seems to be all *gathered* from the zero and the negative side compared to the people that were in the Enrichment program because this is kind of *dispersed* off and this is like, *gathered* in the center. (Jan)

From these excerpts, we can collect a set of terms under the umbrella *spread* that indicate similar notions: *spread out*, *scattered*, *evenly distributed*, *dispersed*. Antonyms include *clumped*, *grouped*, *bunched up*, *clustered*, *gathered*. Three other terms, *concentrated*, *tight*, and *close in*, appeared too infrequently to compare, but conjectured to be similar. Note that many of the respondents' articulation of variation accompanied or integrated distribution shape, argued by Bakker (2004) to be a key entry point for understanding variation. The next section will look at language the respondents used to describe distribution – what we call *distribution chunks*, or context-relevant distribution subsets.

Expressions of distribution: Meaningful chunks

This subsection will tackle the second dimension of nonstandard statistical language used by the prospective teachers in the study: observations regarding the distribution. Although our goal in the study was to investigate the way the prospective teachers articulated notions of *variation*, we found that many of their observations relating to the distribution were also rich with notions of variation. In this subsection, we will discuss three particular kinds of observations of the distribution emerged from the data: triads, modal clumps, and distribution chunks.

Triads. Although describing the data by the number or percentage who improved was very common, not all of the respondents used this criterion to split the distribution into two categories: improvement and non-improvement. Other respondents partitioned each distribution into triads—improving, not improving, and "about the same"—as Maria and Chloe did in January:

Maria: Well the people who weren't in the Enrichment program they did score lower, and the one in the Enrichment program, their scores are kind of varied—some of them did improve, others stayed about the same, and some decreased. … The ones who didn't take [Enrichment] basically stayed the same. There was no real improvement. There's maybe, maybe a few that did, but not so many. While in the Enrichment, there was a lot that did improve. (Jan)

Chloe: Well, uh, it seems like the people that aren't in the Enrichment program, they're staying the same or even getting worse off at, with the next test. But the people in the Enrichment program it looks like, it looks like it's evened out. Like you have some doing better, some doing worse, and some on the same level. There's a little bit more doing worse, but you still got those few that are still doing better. (Jan)

Here, Maria and Chloe partitioned the distribution into three categories: those who improved, those who "stayed about the same" and those whose scores decreased. This seems to indicate that they saw more than just the students who scored above and below an absolute cut point of zero (as described in section 5.1) and that the demarcation between improving and not was more blurred. Their decision to split the distribution into triads ("some doing better, some doing worse, and some on the same level") is consistent with Bakker's (2004) finding that students often naturally divide the distribution into three pieces – low, middle, and high. Viewing the distribution in three pieces is one way that the respondents may have simplified the complexity of the data. This supports work by Kosslyn (1994) who argues that the mind is able to hold no more than about four perceptual units in mind at one time.

Modal clumps. Beyond recognition of the distribution into three chunks–low, middle, and high–several of the respondents focused specifically on the middle portion of the distribution, seeing what Konold et al. (2002) refer to as the *modal clump*:

> To summarize their data, students tended to use a 'modal clump,' a range of data in the heart of a distribution of values. These clumps appear to allow students to express simultaneously what is average and how variable the data are. (p. 1).

A diverse set of expressions indicating awareness of a *modal clump* emerged from several of the respondents' responses:

Chloe: It seems like the most, *the bulk* of them are right at zero. (Jan)

Janet: [The Economically Disadvantaged students] seem pretty evenly distributed. I mean, from this bottom group here, they, the two kids with the highest scores are not Economically Disadvantaged, but that's just a, they don't have that much improvement above the others, above *the main group*. (Jan)

Hope: It's kind of flip-flopped *the big chunk* of this one is on this side, *the big chunk* of this one is on this side. (Jan)

Gabriela: *The majority* of the ones that took Enrichment anyways are still more in the middle. [pause] Or they stayed the same, or they got worse, so I would say that just it's not an effective program. (Jan)

The use of terms like *the bulk of them*, *the main group*, *the big chunk*, and *the majority*, by these respondents all indicate an awareness of a modal clump. In addition, Gabriela's observation that the distributions overlapped factored into her decision about whether she thought the Enrichment program was effective. This suggests that she may have been expecting a more distinct division between the two groups if, in fact, the Enrichment program was working. Gabriela seemed to be basing much of her tendency to split the distribution into three groups by focusing on the *scale* (position relative to zero) rather than on the notion of Bakker's (2004) *low-middle-high*. This perspective changed in her May interview as she added an additional caveat to her description:

> *Gabriela*: If the, if the mean will tell you, that give or, give or take five points [of a drop], that that's 'normal' or 'usual' for them, that then that's not cause to think that 'well, they're not improving' or the program's not working – if they stay around those [negative] five points. (May)

Gabriela's decision about whether the Enrichment program was working changed from her January to May interview based on her interpretation of the middle clump. In January, she indicated that the program wasn't working *because* the middle clump overlapped with the non-Enrichment group, and its location being below zero meant most of the students hadn't improved. In May, however, she argued that because a five-point drop was *typical* for the eighth graders as a whole, you could not use the fact that it was negative to argue against the success of the Enrichment program. Notice that she is using the middle clump here as a description of the average of the group. The notion of this clump persisted for others as well in May and its frequency increased, as evidenced by April:

> *April*: *The majority* of their sample size is on the right side of the mean. Um. And I'd say this is about even, maybe a little more on the opposi-, on the left side of the mean. (May)

Mention of proportion or number improved was very common in the respondents' reports in comparing the improvement of each group (more common than reports of average), and many of the respondents who compared the two groups based on proportion of improvement also saw a *majority* or *main group* clump; this may indicate that modal clumps are strong primitive notions of variation and distribution. The fact that almost all of the respondents used some sort of criterion to separate each distribution (into either two or three groups) also implies that modal clumps may be a useful starting point to motivate a more distributional view.

Distribution chunks. Not all of the respondents who described *distribution chunks* did so with the modal clump or as a way to split the group into low-middle-high. Several of the respondents examined a different meaningful subset of the distribution, most typically a handful of students who improved the most:

> *Anne*: It looks like a few students responded really well to the Enrichment class and improved their scores a lot. … [And] for this, these, this student in particular, but all of these [pointing to a group of high improvers], the Enrichment program worked. I would say that. (Jan)

> *Carmen*: It just seems like the majority of them didn't improve very much. … You still have these way up here—not just the fact that they, that these two improved so much [two highest in Enrichment]—but you have several that went way beyond the average, you know, they went beyond the majority that, of the improvement here. (Jan)

Carmen's descriptions in this early interview indicate she was not only paying attention to whether or not the improvement was above zero, but also made note of the outliers "these two [that] improved so much" and those that "went way beyond the average". In examining *the majority*, and two different groups of high values, she indicates that she is changing her field of view from a fixed partitioning to a more dynamic or fluid perspective. At her May interview, she continued this facility with moving between distribution subsets, but was more specific in her description. A few of the subsets she mentions in this longer exchange are noted (Figure 2):

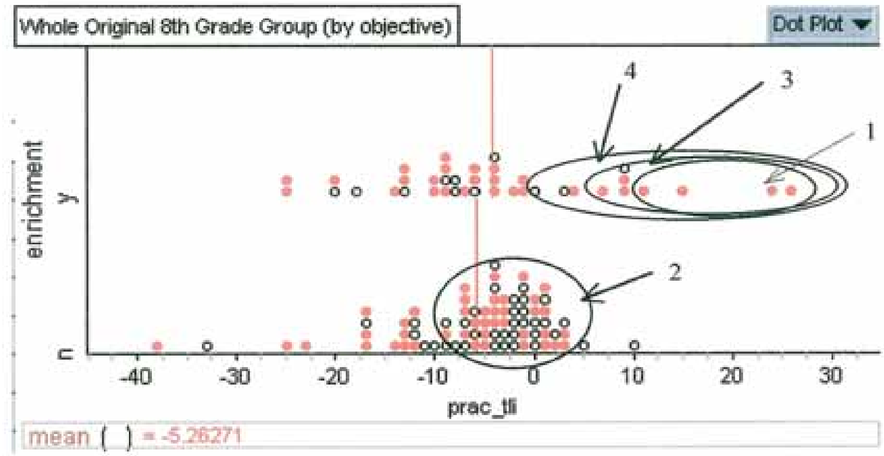

Figure 2. Carmen's description with several meaningful subsets of the distribution circled.

I: Compare the improvement of the students who were in the Enrichment program with the students that weren't.

Carmen: Um, well, … the student with the highest improvement in the Enrichment program was about 16 points above the student with the highest improvement in the regular program. Um. And then, uh, this clump [upper Enrichment], there are others that are higher than the highest improvement, too. I mean there's about four that, um, improved more than the student in the regular program with the highest score [Figure 2, circled portion 1]. Um. It looks like, well there are more people, more students in the non-Enrichment program, and, um, and the non-Enrichment program, um, they also scored considerably lower. I guess the improvement was considerably lower than the students in the Enrichment program. So. I don't know. …

I: Okay. Um. So in your opinion, would you say the program is working? Should they continue it?

Carmen: Um. Well. It seems that it, it is working, that, um. I mean, all of these students improved a lot more than, uh, this big clump of students here [circled portion 2] in the non-Enrichment program, um, but on the same token, it looks like there was about the same that didn't improve in both programs.

I: Do you mean the same number or the same percentage?

Carmen: The same number. Around. Um. … I think, I think it is working and they should probably continue it because of the ones that improved, you know, well there's four that improved much greater than the, than the one that improved ten, by ten in the non-Enrichment group, um, but then there were, there were about eight that improved, uh, more than majority of the non-Enrichment group [circled portion 3]. So even though it wasn't, well, probably one-third, um, showed a considerable improvement [circled portion 4] and I, I would think that that's worth it. (May)

Here, Carmen's ability to interpret and make meaning of the distributions goes beyond recognition of the distribution as the frequency of values of a variable (improvement scores). Besides a quantification of her view of outliers (16 points above the maximum of the non-Enrichment group), one could argue that Carmen is seeing the set of high values in the distribution as more than just individual points, but as a contiguous subset of the distribution. As demonstrated above by the circled areas on the distribution, Carmen was also demonstrating her ability to see several *distribution chunks*, with dynamic borders. This perspective of chunks (as a subset rather than individual points) is more distribution-oriented and indicates that examining chunks, beyond just the minimum and maximum, may be a useful way to encourage teachers to adopt a more distribution-oriented view of the data. Tentatively, the notion of distribution chunks seems to fit somewhere between a focus on individual points and a view of the distribution as a single entity or aggregate (see Konold, Higgins, Russell, & Khalil, under review). As well as having a perspective of distribution as both whole (aggregate) and part (subset), Carmen was also clear in her ability to articulate meanings of both whole and part in terms of the context. For example, she recognized the cut point for "improvement" as values above zero and used the clumps in the distribution to identify the majority of students' improvement scores.

5.3. RELATIONSHIPS BETWEEN CATEGORIES

In this subsection, we briefly examine two perspectives of relationships between and among standard and nonstandard language use, and provide data about the use of language by respondents in January and May. First, we compared the number of standard and nonstandard terms given by each respondent. Recall that the percentage of respondents using standard or nonstandard language changed little during the course (Table 1 and Table 2). However, the mean *number* of distinct terms used by the respondents showed a slightly different pattern from January to May, as shown in Table 3. There was a significant difference between the mean number of standard terms used in May and in January ($t_{16} = 2.84$, $p = 0.01$) and Table 3 shows that the mean number of different standard statistical terms used by respondents increased between January and May. One explanation for this is that the respondents learned (or reviewed) conventional statistical terms during the course. There was no significant difference between January and May in the case of non-standard terms ($t_{16} = 0.77$, $p = 0.46$). In addition, looking at the total numbers of standard and non-standard terms used by respondents in January and May combined, the difference between the mean numbers of terms used was not significant ($t_{16} = 1.73$, $p = 0.10$). That is, in general, respondents likely used no more standard statistical terms than nonstandard statistical observations. Note, however, that there is greater within-group variation for nonstandard terms, suggesting nonstandard statistical language was used less consistently across subjects than standard statistical language.

Table 3. Mean (standard deviation) number of standard and nonstandard terms used by respondents (N=17)

	Standard Observations	Nonstandard Observations
January	2.12 (1.76)	1.94 (2.19)
May	3.88 (1.36)	2.47 (2.07)
Change (from January to May)	1.76 (2.56)	0.53 (2.85)
Total (January and May)	6.00 (1.84)	4.41 (3.16)

Next, we looked at nonstandard language use and whether the dimensions of nonstandard statistical language (variation and distribution) are overlapping or independent. The two dimensions of nonstandard language were suspected to be related. For example, describing a distribution's variation using the term "clumped" also described the distribution as mound-shaped. Because the nonstandard descriptions of variation often contained distribution characteristics, we wondered to what extent articulation of these concepts overlapped. Most of the subjects did not describe the distributions in the same way across interviews. For example, if they mentioned notions of spread in January, it was no more or less likely that they would mention spread in May. We thought it safe, therefore, to combine the types of responses from January and May in order to investigate, as a whole, whether notions of spread and distribution could be considered independent or correlated. Table 4 displays the number of respondents articulating nonstandard expressions of variation and distribution in their interviews. Much of the data in the table falls along the diagonal which implies that most respondents tended to describe either both the variation and distribution in the data or neither. A Fisher's Exact Test (Cramer, 1997) was used to investigate the relationship between these two dimensions and was found to suggest an association between nonstandard expressions of variation and distribution ($p = 0.02$). An interpretation of this could be that the respondents tended to "see" variation and distribution in the data graphs (or not) together. Another possibility is that nonstandard statistical language naturally integrates these concepts.

Table 4. Number of respondents incorporating nonstandard expressions of variation and distribution (N=34)

	Nonstandard Expressions of Variation		
Nonstandard Expressions of Distribution	Yes	No	TOTAL
Yes	13	7	20
No	3	11	14
TOTAL	16	18	34

6. DISCUSSION

The goal of the study was to gain insight into the ways in which the respondents, prospective teachers, expressed notions of variation in comparing data distributions in a relevant context. The task given to the teachers asked them to compare the relative improvement of test scores between two groups of students. Two categories of statistical language emerged from the teachers' descriptions: standard statistical language and nonstandard statistical language. The diversity and richness of their descriptions of variation and distribution demonstrated that the prospective teachers found many ways to discuss these concepts, and that through their nonstandard language, they were able to articulate keen awareness of variation in the data. Two dimensions of nonstandard language were found—observations of spread (variation) and observations of meaningful chunks (distribution). These dimensions overlapped, indicating that either the respondents saw these two notions (or not) together, or that the nonstandard language naturally integrated notions of variation and distribution. In addition, no overall quantitative differences were found between the prospective teachers' use of standard and nonstandard statistical language.

This section will discuss two outcomes of the study: characteristics of nonstandard statistical language or "variation-talk", and elements of the structure of standard and nonstandard statistical language. We will close with a reflection on limitations of the study.

6.1. CHARACTERISTICS OF "VARIATION-TALK"

The subjects in the study expressed concepts of variation using nonstandard language in a variety of ways; we call these ways of articulating variation *variation-talk.* The variation-talk used by these prospective mathematics and science teachers was not so different than the language that emerged in other recent studies of learners' concepts of variation and distribution (Bakker, 2004; Canada, 2004; Hammerman & Rubin, 2004; Reading, 2004), however these studies did not look at nonstandard language systematically and employed other comparison tasks.

The results in this study classified the prospective teachers' *variation-talk* into four types: *spread*, *low-middle-high*, *modal clump*, and *distribution chunks*. The nonstandard language used by the teachers to express *spread*—clustered, clumped, grouped, bunched, gathered, spread out, evenly distributed, scattered, dispersed—all past participles, highlighted their attention to more spatial aspects of the distribution. These terms took on a meaning that implied attention to variation as a characteristic of shape rather than as a measure. This is consistent with Bakker's (2004) description of shape as a *pattern of variability*. In contrast, concepts of variation in conventional statistical language are articulated by terms like *range* or *standard deviation*, both of which are measures.

The other three types of "variation-talk", all nouns, focused on aspects of the variability of the data which required the prospective teachers to partition the distribution and examine a subset, or *chunk* of the distribution. Similar to Hammerman and Rubin's (2004) teachers, the prospective teachers in this study simplified the complexity of the data's variability by partitioning them into bins and comparing slices of data, in this case three slices: *low-middle-high*. Although simple, this is a more variation-oriented perspective than responses that took into account only the proportion or number of students who improved on the test. This approach is consistent with the findings of Bakker (2004), who argued that partitioning distributions into triads may be more intuitive than the more conventional partitioning into four, as in the box plot. The awareness of learners' tendency to simplify the complexity of the data into low-middle-high bins motivated the creators of Tinkerplots (Konold & Miller, 2004) to include the "hat plot" (Konold, 2002) in the software, a representation where users can partition the data into thirds based on the range, the average or standard deviation, the percentage of points, or a visual perspective of the data.

In some excerpts, teachers argued that the overlap of the middle slices was evidence that the Enrichment class was not effective. The focus on the *modal clump* (Konold et al., 2002) in data has been a consistent finding across several studies of both students and teachers, reinforcing both the intuitive nature of seeing variability through slices-in-thirds, and recognizing the potential of using the notion of a modal clump to encourage learners to move from a focus of individual points towards a focus on the *aggregate* of a distribution. Having a lens of a distribution as an aggregate, as opposed to a set of individual points, allows for concepts such as center to be thought of as a characteristic of the distribution rather than as a calculation derived from individual points (Bakker, 2004). In addition, locating a *modal clump* allows the learner to simultaneously express their visualization of the center and spread of the data (Konold et al., 2002), again highlighting its relational nature.

Finally, the use of *other meaningful chunks* by several of the prospective teachers demonstrated their ability to focus on the variability of the data by examining particular subsets of the distribution. This category contained responses as simple as comparing the outliers of one distribution to the "majority" of the other. A more complex visualization of distribution subsets was articulated by Carmen who used several different distribution subsets, with dynamic borders, as evidence for the effectiveness of the Enrichment program. Her facility to fluidly manipulate borders of these subsets highlighted her ability to visualize variation in the data. In addition, each of her *meaningful chunks* was tied back to the context of the problem, indicating that she was able to use them to make meaning of the situation. Although Konold and Bakker have encouraged the conceptualization of a distribution as an aggregate, the articulation of distribution chunks by these teachers suggests that they are thinking of distribution chunks as *mini-aggregates* of the data. Kosslyn (1994) argues that our minds and eyes work together to actively group input into perceptual units that ascribe meaning. We would argue that in articulating subsets of the distribution, the prospective teachers are communicating the perceptual units they are seeing in the data.

The three perspectives of seeing partial distributions—triads, modal clump, and distribution chunks—indicate that there are more than just the two perspectives of distribution that are usually discussed in the literature: single points and aggregate. This third perspective—partial distributions or "mini-aggregates"—deserves further research to investigate the strength of its link to statistical thinking about distributions. This study has highlighted that prospective teachers may use descriptions of partial distributions not only to articulate rich views of variation, but also to use these distribution chunks in meaningful ways that could not be captured using conventional statistical terminology. Even though we classified expressions of variation separately from notions of distribution, this separation was somewhat artificial in that all four types of "variation-talk" expressed a relationship between variation and distribution.

6.2. THE STRUCTURE OF STANDARD AND NONSTANDARD STATISTICAL LANGUAGE

The nonstandard language used by the prospective teachers by its very nature *integrates* the important statistical ideas of variation and distribution, capturing and implying cognitive relationships between notions of center, dispersion, and shape. In contrast, the standard statistical language used by the prospective teachers was by its very nature less relational, with a tendency to express important ideas in statistics as conceptually separate. Terms such as mean, standard deviation, and skewed distribution describe the center, dispersion, and shape of a distribution but they do so in isolation. The overuse of this standard terminology, at least early in learning statistical concepts, may encourage learners to maintain a perspective that statistical concepts are isolated "bits" of knowledge rather than information that can provide insight into relationships in the data. Several difficulties of learning statistics that are documented in research are consistent with a perspective of statistics as isolated facts. For example, seeing data as a set of isolated points rather than developing a "propensity perspective" (Konold, Pollatsek, Well, & Gagnon, 1997), lack of intuition in stochastics (Shaughnessy, 1992), or focusing instruction on calculations, isolated procedures, and graph characteristics in statistics instead of on drawing meaningful conclusions from data (Pfannkuch et al., 2004) may all be reinforced by the *overuse* (too much, too soon) of conventional statistical language.

The process of integrating rather than separating concepts in statistics has been shown to be a productive avenue for developing statistical thinking and reasoning (Konold et al., 2002; Konold & Pollatsek, 2002; Bakker, 2004). This does not imply that one should "teach" nonstandard statistical language as a means to encourage the development of students' intuition about statistics, but rather to encourage their sense-making by acknowledging and encouraging learners' own language. Had we acknowledged only conventional terms in our search for their articulation of variation, we would have lost many opportunities to gain insight into their thinking. For example, only two preservice teachers used the term "standard deviation" in their comparisons, and it can be argued that neither one used this concept to articulate meaning about variation.

An important lesson we learned in trying to categorize the preservice teachers' descriptions was how difficult it was to separate their observations of variation and distribution. For example, when April described the shape of one distribution ("skewed to the left"), she compared its shape with the *spread* of the other distribution ("evenly spread out"). Using the teachers' words, we could see that describing data as *clumped* or *spread out* said as much about the *distribution* of the data as it did about its *variation*. It was not surprising, therefore, when the responses in the two dimensions of nonstandard language (variation and distribution) were found to be correlated.

6.3. LIMITATIONS OF THE STUDY

Although we believe that the results of this study communicate a powerful message about the opportunities of listening to learners' nonstandard statistical language, several limitations of the study must be acknowledged.

- *Population*. The subjects in the study were prospective secondary mathematics and science teachers and the study is not directly generalizable beyond that population. While many of the results may seem to transfer to other groups (for example, practicing teachers), more research would need to be conducted in order to corroborate these results with other populations.
- *Particulars of the task*. The setting may have had strong influences on the responses that the prospective teachers gave during their interviews. Although the data were authentic, the task was not as it did not emerge from the teachers' own desire to know. Therefore, it is possible that the responses given by the prospective teachers may have mirrored an expectation of what they thought the researchers wanted them to say, particularly since the researchers were also their course instructors and the interviews were conducted as part of the course assessment.
- *Particulars of the graphic*. Elements of the graph could elicit responses that may differ with slight modifications. Recall, for example, that the means were marked on the figure and explicitly pointed out to the subjects at the beginning of the task. By drawing specific attention

to the means, we may have been tacitly communicating that the means should have been attended to in addition to or even instead of the prospective teachers' own way of comparing the two groups. The results of the Hammerman and Rubin (2004) study, for example, noted that the teachers in their study were not interested in using means to compare distributions but invented their own ways of making meaning of comparisons in the problems they were discussing.

- *Irregularities in questioning*. In a few of the interviews, the interviewer did not explicitly ask for a decision to be made regarding the effectiveness of the Enrichment program. This meant that comparisons could not be made systematically regarding the prospective teachers' use of evidence towards a decision, which could have elicited richer responses.
- *Generalizations about the course*. Given the unique nature of the course which the prospective teachers undertook, the study was not designed to make generalizations about the effectiveness of such a course in developing learners' statistical reasoning. Many of the compelling elements used in the course and the task were developed from local problems in the implementation of standardized testing. Instead, the study was designed to focus on variation-related language that preservice teachers use at different stages or levels of learning about statistics. Other reports of the larger research study communicate elements of the course that may have had an impact on the prospective teachers' thinking and learning (Confrey et al., 2004; Makar & Confrey, submitted; Makar, 2004).

7. CONCLUSION AND IMPLICATIONS

This study was built on pioneering work by researchers at TERC who first studied teachers' nonstandard statistical language (e.g., Russell & Mokros, 1990), and shows the breadth and depth of nonstandard language used by prospective teachers to describe variation. It follows that teachers need to learn to recognize and value "variation-talk" as a vehicle for students to express meaningful concepts of variation and distribution. This study has contributed to our understanding that preservice teachers majoring in math and science often articulate meaning in statistics through the use of less conventional terminology. Other research studies have shown that school children do so as well (e.g., Bakker, 2004; Reading, 2004; Konold et al., 2002). These studies have begun to articulate rich and productive learning trajectories to move students towards a more distribution-oriented view of data.

Although the preservice teachers in this study were using nonstandard statistical language, the concepts they are discussing are far from simplistic and need to be acknowledged as statistical concepts. Not recognizing nonstandard statistical language can have two pernicious effects. For one, we miss opportunities to gain insight students' statistical thinking. Noss and Hoyles (1996) explain the benefits of attending to students' articulation for creating mathematical meaning: "It is this *articulation* which offers some purchase on what the learner is thinking, and it is in the process of articulation that a learner can create mathematics and simultaneously reveal this act of creation to an observer" (p. 54, italics in original). Gaining insight into student thinking, therefore, is not the only benefit from attending to nonstandard statistical language. Noss and Hoyles also argue that through the process of articulation, students have opportunities to *create* meaning. Sense-making in statistics is an ultimate goal that is often neglected by more traditional learning environments. Valuing the diversity of students' "variation-talk" and listening to student voice (Confrey, 1998) may encourage teachers to shift from the typically procedure-focused statistics courses towards a focus on sense-making. This is an issue of equity if we are to acknowledge the diversity of students' ideas rather than just cover the content and label those who don't talk statistically as being *unable* to do so.

The other problematic effect of neglecting students' nonstandard statistical language is the tacit message that is communicated that statistics can only be understood by those who can use proper statistical talk. Lemke (1990) argues that the formal dialogue of science communicates a mystique of science as being much more complex and difficult than other subjects, requiring that we defer our own ideas to those of 'experts'. By doing so, Lemke argues

> The language of classroom science sets up a pervasive and false opposition between a world of objective, authoritative, impersonal, humorless scientific fact and the ordinary, personal world of human uncertainties, judgments, values, and interests … many of the unwritten rules about how we are supposed to talk science make it seem that way (p. 129).

Although Lemke here is talking about science, the situation in statistics is isomorphic. Countless students complain that statistics (which they call "sadistics") is not understandable to ordinary humans and lacks a connection to sense-making. By communicating statistics as accessible only to experts and geniuses, we are reinforcing a notion of statistics as a gatekeeper to powerful insights about the world, alienating students and denying opportunity to those who do lack the confidence in their ability to make sense of statistics.

Three major benefits come out of teachers' use of their own, nonstandard statistical terminology in describing, interpreting, and comparing distributions. First, they are using words that hold meaning for them and that convey their own conceptions of variation. In the constructivist perspective, knowledge is not conveyed through language but must be abstracted through experience. Nonstandard language carries a subjective flavor that reminds us of this. Through interaction with others, this subjectivity becomes *intersubjective*—one's meaning is not identical to another's, but through further explanation, our meanings become more compatible with the language of our peers (von Glasersfeld, 1991). Second, everyday uses of language are more accessible to a wide variety of students, allowing multiple points of entry to statistical concepts while encouraging teachers to be more sensitive to hearing rich conceptions of variation in students' voice (Confrey, 1998), words that may allow students easier access to class discussions. This is a more equitable, more inclusive stance; one that is contrary to the conception of mathematics (or statistics) as a gatekeeper. Third, if the goal is to provide students with experiences that will provide them with a more distribution-oriented view of data, then nonstandard statistical language that emerges from making meaning of statistical concepts may help to orient students (and their teachers!) towards this perspective. Describing a distribution as "more clumped in the center" conveys a more distribution-oriented perspective in language than stating, say, standard deviation or range to compare its dispersion.

The results of this study are not meant to downplay the importance of using conventional terms and measures in comparing groups. On the contrary, these are very important tools. Our hope is that teachers do not emphasize simply summarizing or reducing the data with conventional measures to make overly simplistic comparisons as we have seen schools do in examining test data (Confrey & Makar, 2005), but rather seek insights into the context the data represent through richer views that include notions of distribution and variation. Further research is needed to gain insight into *how* teachers understand concepts of variation and distribution, as well as to document how teachers support their students' emerging statistical understanding.

ACKNOWLEDGEMENTS

This research was funded by the National (U.S.) Science Foundation under their Collaborative for Excellence in Teacher Preparation program (DUE-9953187). The authors are grateful to NSF for its financial support and to the UTeach program at the University of Texas for their support in this study. The authors wish to thank the organizers of SRTL-3 for their permission and encouragement to make these results available to a broader audience, and also appreciated the feedback offered by our SRTL colleagues. The authors would also like to thank the reviewers who provided constructive, thoughtful, and extensive feedback on earlier drafts of this paper. Insightful and tenacious feedback from Iddo Gal, going beyond his role as editor of SERJ, greatly improved the final version.

REFERENCES

Bakker, A. (2004). *Design research in statistics education: On symbolizing and computer tools.* Published Doctoral dissertation. Utrecht, The Netherlands: Freudenthal Institute. [Online: www.stat.auckland.ac.nz/~iase/publications/dissertations/dissertations.php]

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. 147-168). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Biehler, R. (1997). Student's difficulties in practicing computer-supported data analysis: Some hypothetical generalizations from results of two exploratory studies. In J. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics* (pp. 169-190). Voorburg, The Netherlands: International Statistics Institute.

Boaler, J. (1997). *Experiencing school mathematics: Teaching styles, sex, and setting*. Buckingham: Open University Press.

Canada, D. (2004). *Preservice teachers*' *understanding of variation.* Unpublished doctoral dissertation, Portland State University, Portland, Oregon (USA). [Online: www.stat.auckland.ac.nz/~iase/publications/dissertations/dissertations.php]

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

Confrey, J. (1998). Voice and perspective: Hearing epistemological innovation in students' words. In M. Larochelle, N. Bednarz, & J. Garrison (Eds.), *Constructivism and education* (pp. 104-120). New York, N.Y.: Cambridge University Press.

Confrey, J., & Makar, K. (2002). Developing secondary teachers' statistical inquiry through immersion in high-stakes accountability data. In D. S. Mewborn, P. Sztajn, D. Y. White, H. G. Wiegel, R. L. Bryant, & K. Nooney (Eds.), *Proceedings of the Twenty-fourth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education (PME-NA)*, *Athens, GA,* (Vol. 3, pp. 1267-1279). Columbus, OH: ERIC Clearinghouse, Mathematics and Environment Education.

Confrey, J., & Makar, K. (2005). Critiquing and improving the use of data from high-stakes tests with the aid of dynamic statistics software. In C. Dede, J. Honan, & L. Peters (Eds.), *Scaling up for success: Lessons from technology-based educational improvement* (pp. 198-226). Cambridge, MA: Harvard College of Education.

Confrey, J., Makar, K., & Kazak, S. (2004). Undertaking data analysis of student outcomes as professional development for teachers. *International Reviews on Mathematical Education (Zentralblatt für Didaktik der Mathematik), 36*(1), 32-40.

Cramer, D. (1997). *Basic statistics for social research*. London: Routledge.

Finzer, W. (2001). *Fathom*! (Version 1.16) [Computer Software]. Emeryville, CA: Key Curriculum Press.

Fischbein, E. (1987). *Intuition in mathematics and science: An educational approach.* Dordrecht, The Netherlands: D. Reidel Publishing Company.

Flyvbjerg, B. (2001). *Making social science matter: Why social inquiry fails and how it can succeed again*. Cambridge, UK: Cambridge University Press.

Hammerman, J., & Rubin, A. (2004). Strategies for managing statistical complexity with new software tools. *Statistics Education Research Journal, 3*(2), 17-41. [Online: www.stat.auckland.ac.nz/serj]

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

Hawkins, A. (Ed.). (1990). *Teaching teachers to teach statistics.* Voorburg, The Netherlands: International Statistics Institute.

Konold, C. (2002). "Hat Plots? " Unpublished manuscript, University of Massachusetts, Amherst.

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

Konold, C., Higgins, T., Russell, S. J., & Khalil, K. (submitted). Data seen through different lenses. *Journal for Research in Mathematics Education*.

Konold, C., & Miller, C. (2004). *Tinkerplots* (Version 1.0) [Computer Software]. 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. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics* (pp. 151-167). Voorburg, The Netherlands: International Statistics 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* [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute.

Kosslyn, S. (1994). *Elements of graph design*. New York: W. H. Freeman and Company.

Lemke, J. (1990). *Talking science: Language, learning, and values.* Norwood, NJ: Ablex Publishing.

Makar, K. (2004). *Developing statistical inquiry: Prospective secondary mathematics and science teachers' investigations of equity and fairness through analysis of accountability data.* Unpublished Doctoral dissertation, University of Texas at Austin (USA). [Online: www.stat.auckland.ac.nz/~iase/publications/dissertations/dissertations.php]

Makar, K., & Confrey, J. (submitted). Modeling fairness in student achievement in mathematics using statistical software by preservice secondary teachers. *ICMI Study Volume 14: Applications and modeling in mathematics education*.

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.

Marshall, J., Makar, K., & Kazak, S. (2002). Young urban students' conceptions of data uses, representation, and analysis. In D. S. Mewborn, P. Sztajn, D. Y. White, H. G. Wiegel, R. L. Bryant, & K. Nooney (Eds.), *Proceedings of the Twenty-fourth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education (PME-NA), Athens, GA*, (Vol. 3, pp. 1293-1304). Columbus, OH: ERIC Clearinghouse, Mathematics and Environment Education.

McClain, K. (2002). Supporting teachers' understanding of statistical data analysis: Learning trajectories as tools for change. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics, Cape Town, South Africa* [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute.

Moore, D. (1990). Uncertainty. In L. A. Steen (Ed.), *On the shoulders of giants* (pp. 95-137). Washington, D.C.: Academy Press.

Moore, D., & McCabe, G. (1993). *Introduction to the practice of statistics* (2nd ed.). New York: Freeman.

National Research Council (1999). *High stakes: Testing for tracking, promotion, and graduation*. Heubert, J. P., & Hauser, R. M. (Eds.), Washington, D.C.: National Academy Press.

National Research Council (2000). *Inquiry and the National Science Education Standards: A guide for teaching and learning*. Washington, D.C.: National Academy Press.

National Research Council (2001). *Knowing what students know*. Washington, D.C.: National Academy Press.

Noss, R., & Hoyles, C. (1996). *Windows on mathematical meanings: Learning cultures and computers*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Pfannkuch, M., & Begg, A. (2004). The school statistics curriculum: Statistics and probability education literature review. *Mathematics Curriculum Reference Group Report to the Ministry of Education.* Auckland, New Zealand.

Pfannkuch, M., Budgett, S., Parsonage, R., & Horring, J. (2004). Comparison of data plots: Building a pedagogical framework. Paper presented at the *Tenth International Congress on Mathematics Education (ICME-10)*, Copenhagen, Denmark, 4-11 July.

Reading, C. (2004). Student description of variation while working with weather data. *Statistics Education Research Journal, 3*(2), 84-105. [Online: www.stat.auckland.ac.nz/serj]

Reading, C., & Shaughnessy, J. 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.

Russell, S. J., & Mokros, J. R. (1990). What's typical: Teacher's descriptions of data. Paper presented at the *Annual Meeting of the American Educational Research Association*, Boston, MA.

Schoenfeld, A. (1991). On mathematics as sense-making: An informal attack on the unfortunate divorce of formal and informal mathematics. In J. Voss, D. Perkins, & J. Segal (Eds.), *Informal reasoning and education* (pp. 311-343). Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

Shaughnessy, J. M. (2001). *Draft of grant proposal sent to the National Science Foundation*.

Shaughnessy, J. M., Watson, J., Moritz, J., & Reading, C. (1999, April). School mathematics students' acknowledgement of statistical variation. In C. Maher (Chair), *There's More to Life than Centers*. Presession Research Symposium, 77th Annual National Council of Teachers of Mathematics Conference, San Francisco, CA.

von Glasersfeld, E. (Ed.) (1991). *Radical constructivism in mathematics education.* Dordrecht, the Netherlands: Kluwer Academic Publisher.

Watkins, A. E., Scheaffer, R., & Cobb, G. (2003). *Statistics in action: Understanding a world of data*. Emeryville, CA: Key Curriculum Press.

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

KATIE MAKAR

School of Education, Social Sciences Bldg

University of Queensland

Brisbane, Queensland 4072

Australia