TEACHING AND LEARNING

AN ANALYSIS OF STATISTICAL REASONING

LAURA SASLOW

Thesis under the direction of Richard Lehrer and Leona Schauble

This study analyzes the development of statistical reasoning during several mathematics classes of an intact fourth-grade classroom. The teacher and her students were members of a multi-year teacher-researcher collaborative effort. In all, fourteen class sessions were videotaped, one on January 27, 2000 and the rest from April 6, 2000 to May 18, 2000. Data sources include this video recording, made using a single camera, and rough transcripts of the class talk, written at the time of the videotaping. In the course of the lessons, the students and their teacher worked through statistical ideas and problems about data describing differently sized bubbles, people, and plants. The lessons were analyzed several different ways, including looking at the order and connectivity of turns of talk, the frequency of mention of different topics, the comparisons made between different data sets, and how arguments were formed about expectations and distributions.

Approved by

Richard Lehrer

Leona Schauble

AN ANALYSIS OF STATISTICAL REASONING

By

Laura R. Saslow

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Teaching and Learning

August, 2005

Nashville, Tennessee

Approved by

Richard Lehrer

Leona Schauble

ACKNOWLEDEGMENTS

This work was made possible through the generous support of the Vanderbilt University Dean's Graduate Fellowship and the Peabody Dean's Fellowship.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

This study focused on the development of statistical reasoning in a data-modeling context, with a particular focus on the development of understanding about variation. The analysis followed a fourth-grade class and their teacher as they worked through statistical ideas and problems about data describing differently sized bubbles, people, and plants. I explored how the work of teaching and learning unfolded and was accomplished in the contexts and interactions of the classroom, analyzing the teaching and learning of distribution and variation in a series of mathematics lessons.

This analysis had several goals. One was to document the ways in which students came to statistical understandings, namely the growth and development of their reasoning about variation and distribution in a classroom context in which statistical description could give students new and different views of natural phenomena. Another goal was to conduct comparative analysis of different methodologies for the study of interaction for the "same" episodes. Although the raw data were the same for each of the analyses, different aspects of the lesson were highlighted, and therefore different conclusions could be made.

To conceive of the world statistically, one must be able to generate, test, and revise models of the world (Lehrer & Romberg, 1996). It follows, then, that when learning about statistics, students should be put in the position of inventing and revising models in a real world context. As Cobb and Moore (1997) discuss, "Statistics requires

a different kind of thinking, because data are not just numbers, they are numbers with a context" (p. 801). The National Council of Teachers of Mathematics (NCTM, *Principles and Standards for School Mathematics*, 2000) advises that middle school students "should be driven by a desire to answer questions on the basis of data. In the process, they should make observations, inferences, and conjectures, and develop new questions" (p.251-252). Students must understand the context as well as the statistics themselves in order to truly think statistically. The context influences what and how one chooses to measure, how one interprets what is most relevant, and how to best structure and interpret the data.

In this study, the context included the natural variation inherent in differently sized bubbles, people, and Wisconsin Fast Plants™. It was hoped that the students would develop, critique, and revise data-based arguments about the nature of growth, variation, and differences between data sets about phenomena that they found personally intriguing. It was also a goal that the students would learn to see the data sets as distributions. The particular students in this study had a strong background in geometry and measurement, so it was conjectured that the students would drawn on their understanding of geometry and measurement in order to understand variation and distribution.

According to Konold and Khalil (2003), typical standardized tests of statistical understanding focus mainly on lower-level encode/decode skills such as being able to read a graph of statistical data or create a display from raw data. They suggest that students should instead be able to compare and contrast two sets of data, understand the relationship between two variables, and be aware that increasing sample size stabilizes

measures of the group characteristics. It follows, then, that it would be important to examine what sorts of mathematical thinking the students in the fourth-grade class of this study were doing. Were they simply encoding or decoding the graphs or were they comparing and contrasting the data in more nuanced ways? Did they develop an understanding of group characteristics? What sort of evidence would count as an example of these different ways of thinking about the data?

Several factors are important when trying to make inferences about differences between two distributions. Some of those factors include the spread of the data (such as the standard deviation and interquartile range) and the shape of the data (such as whether or not it is unimodal or bimodal). The most important aspect of any of these factors is that students are able to see the trend or distribution of the data, not just individual points of data or collections of data points in an interval. Distribution was not expected to be an obvious way to think about the variation of the data. When, for example, does one run into the idea of distribution in the real world? In addition, to think about distribution requires one to go beyond individual cases to groups of cases, or populations (Mayr, 1998). It can also be difficult to see trends because data can be very variable and show plenty of exceptions to the trend (Konold & Khalil, 2003).

Not only was it hoped that students would come to understand the differences between two different distributions, but that they would see that the data was part of a stochastic process. Thus, in order to guide the students to think about the data in this way, sampling was developed in different contexts. It was hoped that students would be able to see the differences between samples of data and to make educated guesses about the population from which the samples were drawn. As Konold and Khalil (2003)

suggest, inferring about a population from a sample are the building blocks of formal statistical tests of population inference. Another goal was that students would see the data's distributions as revealing relevant information about the underlying phenomena that it modeled rather than simply describing the spread, shape, or trends. For example, did the students move beyond description to an understanding of what that description meant in terms of the variation and distribution of the data? Did students reason about the variation in the distribution or did they find ways to reduce the variability in order to describe it?

Prior research (Gal, Rothchild, & Wagner, 1990; Konold, Pollatsek, Well, & Gagnon, 1997; Watson & Morits, 1999) has shown that when students compare two different sets of data, three methods are likely. Students may compare slices of data, for example, saying that blue bubbles are better than pink bubbles because blue bubbles have six 30 cm bubbles but pink bubbles only have one bubble that is 30 cm. In other words, the students focus only on one small part of the distributions. Or, students look mostly at the extremes of the data, saying, for example, that blue bubbles are better than pink bubbles because the three biggest bubbles were blue. Alternatively, students might look at a relative cut-off point in the data, saying that blue bubble are better than pink bubbles because 17 of blue bubbles are 20 cm or more but only 9 of the pink bubbles are 20 cm or more. More sophisticated than simply comparing numbers of cases, though, is being able to compare proportions of cases. Instead of noting 17 bubbles are 20 cm or more, students might note that half of the bubbles were 20 cm or more. This sort of reasoning is especially helpful when examining unequally sized data sets. According to this research then, it would be important to pay attention to the approaches the students

took in analyzing the data. Did they look at slices or extreme values? Did they move from counting bubbles to creating proportions? What sorts of comparisons did the students find to be relevant?

CHAPTER II

METHOD

Participants

Participants were an intact class of fourth-graders and their teacher. The teacher and her students were members of a multi-year teacher-researcher collaborative effort in a rapidly growing school district located near a mid-size city in the upper Midwest. This teacher-researcher collaborative focused on mathematics and science education, understanding student learning, and building a community focused around that learning. Details on the theoretical motivation and history of this researcher-teacher collaboration are available elsewhere (Lehrer & Schauble, 2000, in press), as is the teachers' published work (Lehrer & Schauble, 2002).

This teacher taught her fourth-grade students for nearly the entire school day, including subjects other than math. For the study, however, the teacher and her class were videotaped during the mathematics lessons only. In all, fourteen class sessions were videotaped, one on January 27, 2000 and the rest from April 6, 2000 to May 18, 2000. Data sources included a video recording using a single camera and rough transcripts of the class talk, written at the time of the videotaping.

Design of the Study

The investigation took the general form of a design study (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003). As Jim Gee noted in his talk in the spring of 2004

with graduate students of the Teaching and Learning program at Vanderbilt University, classrooms are too messy and filled with too many unknowns for controlled experiments to be very useful. This is because a true controlled experiment is not possible when so many factors cannot be proscribed. Instead, Dr. Gee said that educational researchers should follow the example of physicists who, when faced with a similarly messy and unpredictable system, design models of the system and run multiple trials of data, starting with different initial variables. They do this over and over, creating an iterative process of simulations. They then piece together how initial variables might have influenced the outcomes of the simulations. This allows for patterns and connections to be seen between variables that otherwise might not have been apparent. The study of classrooms might take a similar tack. Researchers could run multiple simulations, study the outcomes, and then try to see what patterns and connections might be made. This paper could be considered a study of an early iteration of using statistics with fourth-graders. The better this iteration can be understood, perhaps the more successful, or at least informed, the next iteration can be. This study did not look at a typical classroom but what happened in a well supported but not irreproducible classroom situation.

Not all iterations can reveal the answers to every reasonable question of that iteration. For example, data collection greatly influences such possible inquiries. It should be noted therefore that several things about this particular classroom and the data set made certain questions impossible. The video was created using only one camera that generally followed the teacher. The camera sometimes panned to the students, especially when they spoke, but a continuous stream of video of the teacher and students was not recorded. This prohibited the study of visual cues, such as gesture, at least in

any systematic way. Most of the time, the person who spoke was identifiable, but this was not always the case, so a study of one particular student could not be easily done either. The analysis of representations was somewhat curtailed as well, because the students examined frequency charts that had been made for them by the teacher. This meant that much of the representation had not been made complicated for the students and was not interrogated by them.

Certain aspects of the classroom lent themselves to analysis. Because a continuous audio record of the classroom was recorded (as part of the video record), the classroom speech was an easy target for analysis. In addition, in the classroom interactions, very few student-to-student speech interchanges (at least those made audible to the whole class) took place. Thus, the analysis had to be very teacher focused and lent itself to questions about the nature and functions of the teacher's revoicing, or how the teacher rephrased student speech and framed ideas. In addition, unlike many fourth-grade students, the students in this particular classroom had studied geometry and measurement. It was conjectured by the original researchers of this study, Dr. Leona Schauble and Dr. Richard Lehrer, that the students would drawn on their understanding of geometry and measurement in order to understand growth and distribution. The data lent itself to questions about whether these students were facile with fractions, and if so, if this ability translated into an understanding of distribution. The classroom activities played a prominent role in the unfolding of events over the days, so an overview of those activities and their relationship to possible student understanding also lent itself to analysis.

In order to do this analysis, as suggested in Erickson and Schultz (1977/1997), I went over the data several times, each time going deeper and deeper into the activities, participation structures, and contexts. At first, I was not sure which of the classes would be most interesting, so I read and re-read, watched and re-watched all of them. The data went from confusing to a clearer pattern of activity. It was especially helpful to go back and forth between reading research about this sort of analysis to studying the raw data again. Each time I reexamined the data, I was able to see more nuances and have more questions to ask.

In the end, I decided to analyze the classrooms using a few different approaches. The hope was that a multi-layered analysis will reveal more than one approach alone would have. First, the approach of Strom et al (2001), in which the researchers traced the mathematical arguments of students in a classroom, was attempted. This analysis is a very micro, small-scale one that relied heavily on accurate coding. The approach draws somewhat from cognitive science, in that the method had as an underlying belief that the class discussion traced through a problem space. This approach was used several different times, each time coding different aspects of the discussion and problem space.

The second type of analysis focused more closely on how the students dealt with expectations of future experiments, with an eye towards how this thread of talk might help reveal what these students understood of the statistics and to trace how this thinking changed over time.

CHAPTER III

RESULTS

Analysis of Turns of Talk

The first approach analyzed coded turns of talk. This treated the classroom discussion as a path that moves through a problem space over time. This analysis was performed primarily on the class session that took place on January 27th. During this class session, the students compared two frequency charts, one with the diameters of 63 bubbles blown with blue bubble solution and 63 bubbles blown with "super" blue bubble solution. An implicit goal of the lesson was to try to decide which of the bubble solutions had blown "better" bubbles. During the discussion, the students compared the distributions using extreme values, typical numbers, ranges, fractions of the total for each mode, fractions of bubbles above the same cut-off point on each graph, medians, and fractions above the median of one graph to that same spot on the other graph.

Order and Connectivity

In order to do the micro analysis on turns of talk of this class session, I tweaked Strom et al.'s (2001) approach somewhat to highlight the role of the teacher. Also, to map the activities and ideas onto my fourth-grade lesson, which in terms of content was quite different to Strom et al.'s lesson, required quite a bit of changing. I created codes for the teacher's remarks that were rephrasings of student speech (13 codes), conclusions that class made about the information (6 codes), procedures that the class used to

analyze the information (18 codes), and prior history relevant to the information (3 codes). As this was the first try at this analysis, these codes were based loosely upon the codes used in Strom et al.'s paper and upon the discussion types that on first pass seemed most central to the discussion. This coding highlighted the role of the teacher, the steps the students took in their analysis, and which analyses the students thought were significant. Please see Appendix A for more detail about these codes. I then coded all 138 turns of talk that were present during the lesson and created an adjacency matrix (Table 1) which revealed which types of talk were followed by other sorts of talk.

Table 1: Adjacency matrix for the lesson collapsed by semantic category (both teacher and student initiated speech turns)

Initial / terminal ↓ / →	To (Q)	To (C)	To (P)	To (G)	Out-Degree
From History (Q)	2	0	4	2	8
From Conclusions/Concepts (C)	0	1	5	3	9
From Procedure (P)	6	1	30	31	68
From Teacher Revoicing, etc. (G)	2	9	26	18	55
In-Degree	10	11	65	54	

I also made a chart of the information in Figure 1.

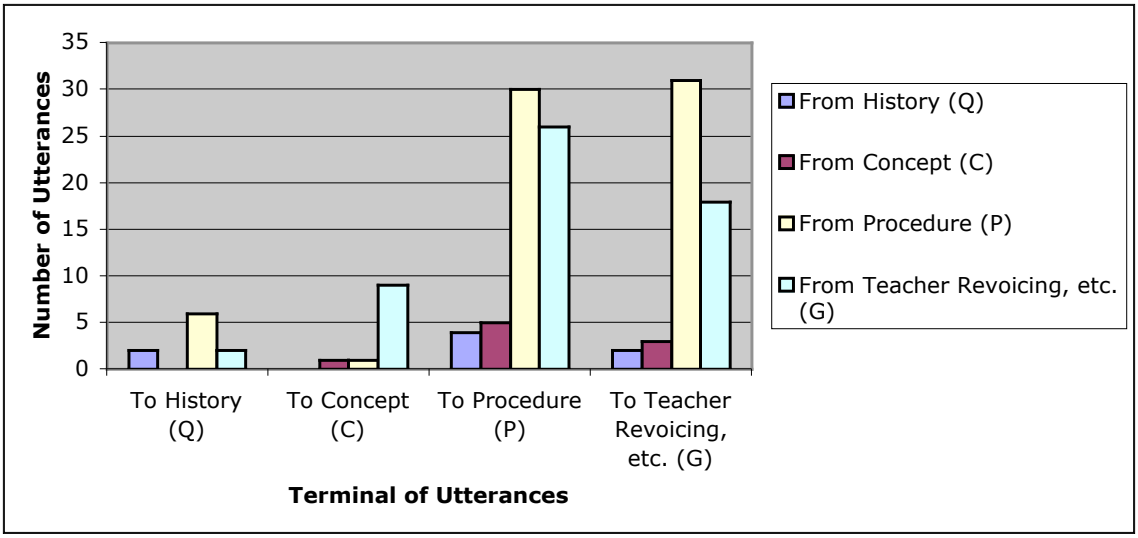

Figure 1: Graph of adjacency matrix for lesson collapsed by semantic category (both teacher and student initiated speech turns)

What this matrix analysis reveals is that most of the concepts/conclusions about the distributions were said after the teacher had revoiced the students' ideas back to them. Perhaps, then, the teacher's revoicing of ideas was an important part of helping students to understand. Instead of having many ideas float around the discussion, the teacher, by repeating and reframing ideas, may have helped focus the students' thoughts to a point where they were able to build more fully upon the ideas. This lends credence to the idea that the teacher's role in the classroom talk was central; students generally found their ideas to lead to conclusions after hearing the teacher revoice those ideas.

In addition, the evidence seemed to show that teacher revoicing and procedures were intertwined in the conversation and central to it. For example, teacher revoicing was often preceded by procedures and or other teacher revoicing. Procedures were generally preceded by other procedures or by teacher revoicing.

The first party to mention a topic, be it the teacher or a student, was also noted. The teacher first mentioned comparing the ends of the range, comparing the whole range, comparing mode by count, comparing mode by proportion, comparing proportions and fractions past a cutoff point on each graph, comparing by count past a cutoff point on each graph. The students mentioned all of the other codes of talk first. Thus, although the teacher played a central role in the talk, she did not mention all of the ideas first.

The procedures that the students tended to bring up first, however, mostly included noticing one area of a graph such as the median, the center of a cluster, or a particular value. This may mean that the students were more inclined to look at the slices of the data (which the research suggested would be a likely first step) and that the teacher was need to help them think more about count and proportion.

The students were the first to mention the making of a fraction, however, and they had more turns of talk about creating and comparing fractions than the teacher. In this context, fractions were used to describe parts of a distribution. Thus, instead of saying that 5 bubbles were a certain size or greater, the students would use fractions to say, for example, that 1/3rd of the bubbles were a certain size or greater.

Code centralities were also examined. The centrality of each code was found by summing the distinct codes that connected to and from particular nodes. The top eight codes, in terms of centrality were P11 (creating proportions or fractions), G8 (teacher revoicing of technique or procedure), G11 (teacher revoicing in which the teachers asks if ideas are sufficient to answer the guiding question according to standards of discussion or asks about meaning of an idea), P13 (comparing proportions or fractions

past cutoff points on each graph), P0 (students' short responses to teacher question about technique), G2 (teacher revoicing about the standards of discussion), P2 (comparing the ends of the range), and G4 (teacher revoicing in which the teacher asks for clarification of a technique or procedure). Half of the top eight central nodes were teacher revoicing, again showing the centrality of the teacher during this classroom discussion. Her speech was highly tied to other types of speech. Also, the students' comparing and creating fractions was highly tied to other types of speech, which reflects how fractions played a central role in the mathematical argument. Comparing proportions past cutoff points and comparing the ends of ranges were also highly central to the argument, which reflects how these aspects of the argument were highly connected to other ideas. (Comparing the proportion past cutoff points was one of the main way the students found as a way to make a sufficient argument about which bubbles were better.)

Discussion Type by Sections of Time

Another analysis was done on the turns of talk during the January 27th class session. It looked at both teacher and student utterances, combined, and examined the prevalence of different types of codes in six different ten-minute sections of the class. The result is shown in Figure 2.

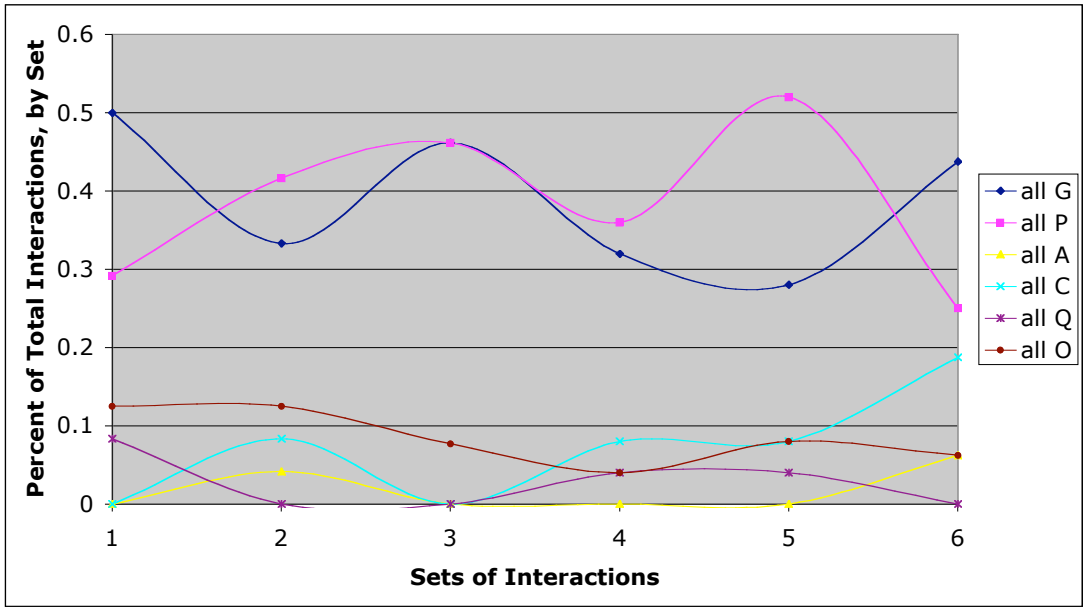

Figure 2: Student and teacher initiations, by type, compared over ten-minute sections

For this analysis O (basic operations) and A (alternative hypotheses) were coded separately, from G (teacher revoicing), P (procedure), C (conclusions) and Q (history and context). The results show that G (teacher revoicing) and P (procedures) were highly prevalent. A (alternative hypotheses, all suggested by students) was present in the second and sixth sections of the argument. C (conclusions/concepts) was mentioned in the second, fourth, fifth, and the sixth ten-minute sections. Q (history and context) was mentioned in sections one, four, and five of the argument, while O (basic operations) was mentioned in all sections.

What this analysis may show is that the argument seemed to be sustained by procedure by the students (and a steady but small background of basic operations) and teacher revoicing.

Frequency of Mention

A third analysis of turns of talk during the January 27th class session was performed, this time collapsing the categories and placing an emphasis on frequency of mention, not order or connectivity. Teacher revoicing (3 codes) and procedure (32 codes) were now of primary emphasis. The teacher revoicing codes became less relevant in this analysis, because most of the teacher revoicing was labeled as procedure, at least if any procedural content was present. Thus a turn of speech that had been coded as a rephrasing of procedure during the first analysis became coded with whatever that procedure was that the teacher was rephrasing in the second analysis. In this way the content, be it from the teacher or the students, became emphasized. The procedural codes now recognized if the comment was noting something only in one graph, comparing generally, comparing by counting, comparing by creating a proportion, or comparing by shape of the distribution. Please see Appendix B for more detail about the coding.

The five most prevalent procedural codes overall in this analysis were N19 (creating and comparing fractions, mentioned more by students that the teacher), N12C (comparing between two cutoff points by proportion, mentioned equally by the teacher and the students), N4 (comparing whole range, mentioned equally), and N8C (comparing modes by proportion, mentioned more by the teacher). This analysis corroborates other evidence that the students talk quite a bit about fraction creation and comparison, in order to help them compare the distributions.

When the codes were combined, regardless of whether the discussion was general, about count, proportion, or shape, the most prevalent features of the discussion (at least by turns of talk), were N12 (comparing between cutoff points, teacher dominated), N8 (comparing modes, teacher dominated), N19 (creating and comparing fractions, student dominated), and N4 (comparing whole range, balanced). This again shows the dominance of the teacher in discussion and the relevance of these particular procedures. The teacher dominated half of the top four discussion types. This analysis does not help reveal the goals of the discussion particularly well. Instead it helps to show the path of the discussion.

The codes were then combined a different way, this time by whether the discussion was general, by count, proportion, etc. The results can be seen in the figure on the next page, Figure 3.

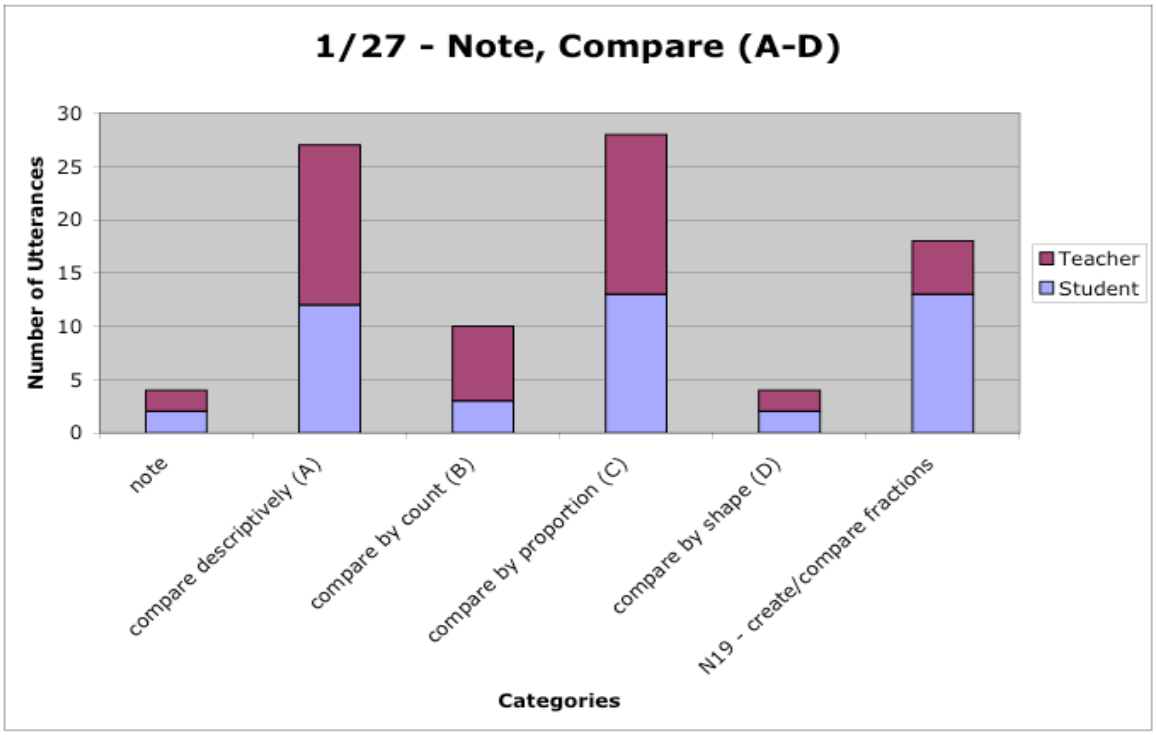

Figure 3: Analysis of discussion on January 27th, by discussion type and speaker

The descriptive comparisons and comparisons by proportion were most prevalent, followed by the creating and comparing of fractions and then comparing by count. It is interesting to note that the comparing by count was dominated by the teacher but the students dominated the create and compare fractions discussion, perhaps showing that the students did not seem to place too much emphasis on comparing count, but once the numbers were available, the were able to play with them in ways that let them make fractions. This might be significant because it suggests that they were comparing ratios, not counts. Although the teacher somewhat dominated the compare

by proportion discussion, this was perhaps off-set by their dominance in the create and compare fraction talk. This analysis may show that the teacher was especially needed to point out to students that the comparisons could be done by counts of various parts of the distribution.

Alternatively, after examining who, whether the teacher or the students, commented about a particular topic, one might be able to create an argument that the teacher pushed the students towards thinking about count. Then the students made and compared fractions from the count (which one might label as an application of part/whole reasoning). By reducing the distribution to numbers that they could compare (be they modes, medians, counts, fractions, etc.) the students effectively reduced the variability in the distribution. This may have delimited the possibility that they would reason about the distribution as a whole or the variability within it.

Comparisons Made

In this analysis the frequency and connectivity of the turns of talk were ignored. Instead, the actual comparisons made between the two graphs were highlighted in order to try to find when it was that students thought that they had made a comparison that could really tell them something meaningful about the two graphs. This analysis showed that small differences in count were interpreted as inconsequential. Three different fraction comparisons were made: 1/9 to 1/6, 1/6 to 1/3, and 1/6 to 1/2. The last comparison was the most consequential for answering the question, according to the class, and it was also the comparison with the largest difference between the fractions. This analysis highlights the underlying math behind the discussion, and shows how the

students' ability to create and compare fractions was very relevant to their ability to be able to find comparisons to be meaningful and consequential.

A Trace of the Argument about Noting and Comparisons, and Expectations

In the next analysis, I looked more closely at eight categories, noting something, comparing descriptively, comparing by count, creating and comparing fractions, comparing by proportion, comparing by shape, and talking about expectations of typicality. In this analysis, comparison by shape was more rigorously coded: only talk that discussed the underlying statistical meaning of the shape was coded as such. Commenting on plateaus or mountains in the data would now be coded as descriptive comparison. Next, I traced the argument through these categories of talk, for 6 different sections, each representing ten minutes of talk. I also created bar graphs of each of the six sections of talk. An example of such a bar graph is in Figure 4.

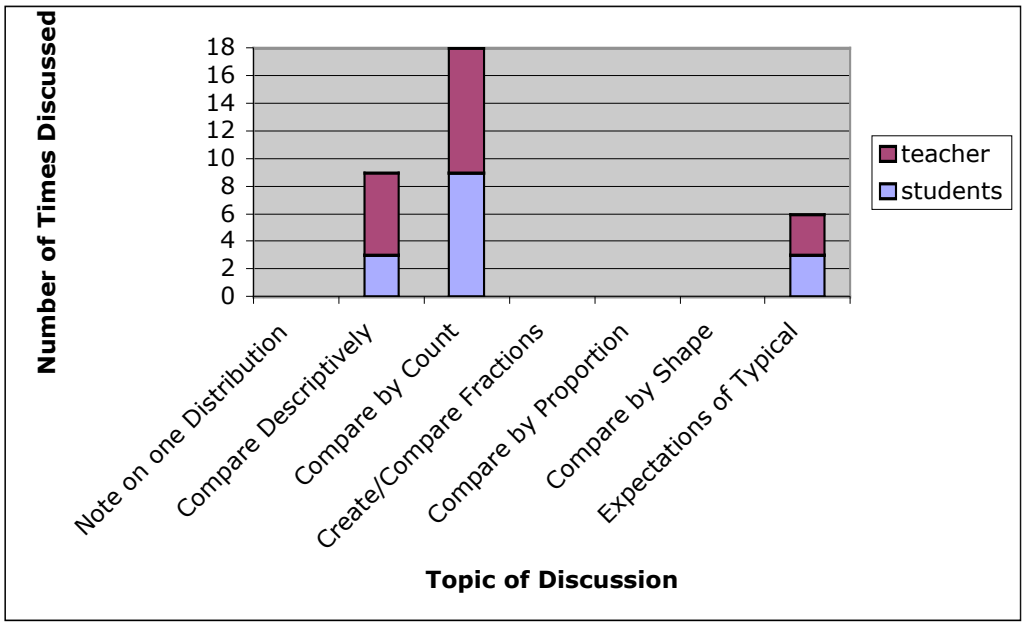

Figure 4: Graph of second ten-minute section of classroom talk, January 27th

The first 10-minute section of the class had few turns of talk that were based around content. The turns of talk that did fall into one of these categories were noting on one distribution, comparing descriptively, and comparing by count. The second 10-minute section had some talk about expectations of typicality (the only section have this category), in addition to descriptive and count comparisons. The third 10-minute section added talk about creating and comparing fractions and comparing by proportion. The fourth and fifth sections had the same categories, but comparing descriptively tapered off. By the last section of talk, the teachers and students mentioned only comparing and creating fractions and comparing by proportion. None of the turns of talk had comparison of shape according to the newer, more rigorous standards.

This analysis shows that the class did generally trace through a problem-space that started with noting something on a description, comparing something descriptively between the two distributions, comparing by count, and using fractions to compare by proportion. One reason for the lack of talk about shape may be that shape was taken as shared—the representation was generated conventionally, and so, different senses of shape were not contested.

A Trace of the Argument about Distribution

In order to focus more specifically upon how the students examined the distribution, another revised version of the coding was created. This time, the differences between noting a points of data, sections of the data, or the overall patterns of the data were taken into account in the hopes that this progression would better characterize increasing sophistication in thinking about the distribution and variation.

For the category of points of data, sample types of data that could be noted could be extreme values, modes, “towers” (numbers with a high frequency with numbers with low frequency adjacent to them), mid-ranges (halfway between the highest and lowest recorded number for a data set), mid-chunks (points in the middle area of a group of high frequency data points), and medians (the middle number found after ordering the data set from lowest to highest).

For sections of data, sample types of data that could be noted could be ranges, extreme value cut-point sections (sections of the data above an extreme data point), mode cut-point sections (sections of the data above the mode), empty bins (data with zero frequency), median cut-point sections (sections of the data above the median), clumps (sections of the data with high frequency data points adjacent to one another), and mid-fifty percent cut-point sections (sections of the data covering the middle fifty percent of data points).

For the overall pattern of the data, sample types of data that could be noted could be the overall pattern around modes, ranges, empty bins, and clumps.

When these aspects of the distribution are noted, several different approaches could be used. An aspect could be pinpointed, i.e. where an aspect is located in the data set could be noted. An aspect might also be counted, that count could be turned into a fraction or a percentage, and then those fractions or percentages could be compared. The density of an aspect in the distribution could be discussed as could where in the distribution an aspect is located and expectations about how an aspect might replicate. For example, a student might pinpoint the mode, count a mode cut-point section, create a fraction of the total distribution that the mode cut-point section represents, and then

compare that fraction to another fraction from a different cut-point section of a different distribution.

Much of the work of the students occurred in the noting of an aspect in one distribution and in comparing distributions. A possible hierarchy of comparisons, from least to most complicated, might be comparing the distributions by count, by creating and then comparing fractions, by descriptive comparisons, and then by shape of the distributions. Other student talk revolved around expectations, alternative hypotheses, and deciding if the information was sufficient to answer whatever question was being asked about the data.

Analysis of Expectations

The previous approaches emphasized individual classes, not the sequence of instruction. This has the weakness that it does not give an overview of how thinking may have changed over time. A different approach was taken in order to see how one particular area of statistical thinking, expectations about sampling and repeating experiments, altered across time. I built upon the previous work and created codes with which to analyze the turns of talk. Then, I noted areas of talk when the class discussed expectations and attempted to use the codes to mine the talk.

I decided to examine whether the classroom talk distinguished points of the data, sections of it, or the overall pattern. Similar to the final analysis of the previous section, for points of the data, the different things noted might be, from possibly least complicated to most complicated: extreme values, modes (or towers), mid-ranges, mid-chunks, or medians. For sections of the data, the different things noted might be, from

possibly least complicated to most complicated: ranges, extreme value cut-point sections, mode cut-point sections, empty bins, median cut-point sections, clumps, and mid-fifty percent cut-point sections. Other distinctions include how the students discuss these aspects, such as whether they pinpoint (locate something in the frequency chart), count, create fractions, compare fractions, note the density of the information, or locate something relative to the entire distribution.

It was conjectured that students would have a variety of anticipations about their expectations for a repeated experiment or for what was typical or normal. Some options for expectations might be that students would expect truncated tails of the distribution or that they would be skeptical about the likelihood of resampling low frequency data points. It was also expected that fractions would be used as a tool for expectations, such as using fractions to understand the chances that a particular data point would be resampled. A discussion of several class's talk about expectations follows.

On January 27th, the first class videotaped for this study, students were asked to guess about what the resulting frequency graph would look like if a data set of 63 bubbles were to have half of the bubbles sampled. Two of the students reasoned about the likelihood of sampling the mode of 30 again, believing that one would likely get around half of the 30s but that it might not still be the mode. One of the students reasoned using fractions, commenting that since 1/6th of all measurements are 30, one would have a half of a sixth of a chance of getting 30 in the sample. By focusing on mode, these students focused on a slice of data. Next, another student discussed the range of the sampled data, saying that he doubted that the range would be as long. In other words, this student expects sampling to lead to truncated tails of the distribution.

Later, that same student commented that the shape of the sample would be different and that the "little bumps," or the numbers with a frequency of one, would be unlikely to be there in the next sample. Similarly, a different student commented that the numbers with a frequency of one were unlikely to be sampled again. Another student commented that there would be a 1/63 chance of getting an outlying number.

On April 6th, the next class session videotaped for the study, the teacher asked the students to think about what most of the blue bubbles they might blow would be sized. One student suggested that she would expect the bubbles she blew to be around 23 or 24 cm because those measurements were in the middle of a chunk (cluster, collection) of data. Another student suggested from 18 to 21 because that was were she thought the cluster was. A different pupil disagreed, saying that one should expect bubbles between 21 and 25 because that was the middle of a big chunk, and it encompassed 10 bubbles, which was a sixth of all of the bubble measurements. The teacher then asked about the other 5/6th of the bubbles, pushing the students to think more deeply about what most of the bubbles should mean. Later, another student remarked that perhaps one could expect bubbles from 18 to 30. The teacher reminded the class that from 18 to 30 was 26 bubbles or about a third of the bubbles. She then asked if a range that encompassed a third of the measurements gave someone a better idea of the sorts of measurements that they might get (as compared to, for example, expectations that covered only a sixth of the bubbles). The class agreed that a third seemed like a reasonable estimate.

In the April 6th class the students therefore used clusters and fractions to help them describe what was normal or typical for the bubbles they were discussing. The

students seemed to show a preference for choosing ranges that encompassed clusters of data, especially when no other clusters existed. The students also seemed to show preference for mid-clusters. Then, once sections of data had been chosen, the class used their knowledge of fractions in order to pick ranges that encompassed enough of the entire distribution of data in order to be normal or usual. In this case, the class agreed that a third was sufficient. This is an interesting quantification of what a mid-cluster might mean.

The next day, on April 7th, the students grappled with the question of what would happen to the range of a frequency chart of their heights if instead of being measured in centimeters their heights had been measured in inches. One student suggested that the range would be smaller if the heights had been measured in inches because the measurements that were 44 to 46 centimeters would all be on the same inch. When it was noted that 11 bubbles were in the middle group and 10 were scattered elsewhere, two students said that they thought that this pattern of clumping and scattering would not changed if the measurement unit of the heights were to change. When the teacher asked how the shape of the frequency chart would change, a student suggested that the measurements would be more clumped together and that the range would be smaller. The discussion from this class showed that some of the students began to think about the apparent density of measures and how it might change with the unit of measure, but that this sort of thinking was a struggle.

On April 11th, the class was videotaped, but due to technical problems, this videotape is no longer available. Notes taken at the time of videotaping will serve as somewhat of a substitute for this day. The teacher asked what would happen if ten data

points were sampled from the data set of the students' heights in inches. Specifically, she asked what number they would find the most of and what would be normal for the new, sampled data set. One student suggested a range of 55-57 inches. Another suggested 54-58 inches. Both of these suggestions described areas around the main cluster of the data. A student noted that only two outliers existed, so that with such a small sample, it would be unlikely that all of the people in the new population would be above the median of 56 inches. The teacher asked if 56 inches was a typical height for the whole 4th grade. A few students said that they expected a bigger range. Then, a student asked what the shape would be for the new data.

The next day, on April 12th, during a discussion comparing two different data sets of bubble data, a student asked what would happen if another class of students wanted to know what would be a normal size for the pink bubbles that the class was examining. Unlike all of the other expectation comments, this one was student-initiated. The student went on to comment that because the pink bubble set had few gaps in the data, it would be hard to say what was normal. She thought that normal would have to be a relatively large section of the data, such as from 10-30 centimeters. For the other data set of blue bubbles on the other hand, she said that one could say normal was in a smaller section of the data, such as 24-30 centimeters. This comment suggests that the student understood that a flat, highly variable distribution has a very wide range of what would be a normal or expected value. It is interesting that the comment about expectations was initiated, in that it may mean that the idea of expectations of normal had become an accepted, common, or at least possible way of thinking about distributions.

CHAPTER IV

DISCUSSION

There were two goals of the analysis, documenting the ways in which students came to statistical understandings and conducting comparative analysis using different methodologies for the study of interaction for the "same" episodes. The various analyses concentrated on different indexes of what statistical understandings might be, varied in terms of the scale of their analysis, and thus were able to find different meaning out the class sessions.

These different sets of analyses uncovered various aspects of the class session in January. For example, the teacher was very important to the discussions. Her revoicing led to student conclusions and led students to create fractions from the counts of data. The class's discussion generally went in the order of noting something on a description, comparing something descriptively between the two distributions, comparing by count, and using fractions to compare by proportion. Comparing by count was dominated by the teacher but the students dominated the create and compare fractions discussion. In terms of centrality, teacher revoicing, the students' comparing and creating fractions, and comparing proportions past cutoff points were highly central to the discussion.

This reflects how fractions and comparing the proportion past cutoff points were two of the main ways the students found to make a sufficient argument about which bubbles were better. For example, another analysis showed that three different fraction comparisons of sections of the data were made during the discussion: 1/9 to 1/6, 1/6 to

1/3, and 1/6 to 1/2. The last comparison was the most consequential for answering the question, according to the class, and it was also the comparison with the largest difference between the fractions. This highlights that the students' ability to create and compare fractions was very relevant to their ability to be able to find comparisons to be meaningful and consequential.

The procedures that the students tended to bring up first mostly included noticing one area of a graph such as the median, the center of a cluster, or a particular value. This may mean that the students were more inclined to look at the slices of the data (which the research suggested would be a likely first step) and that the teacher was need to help them think more about count and proportion.

In the last, brief analysis of expectations, it was shown that students used clusters of data points and fractions to help them describe what was normal or typical for the bubbles they were discussing. The students seemed to show a preference for choosing mid-clusters and ranges that encompassed clusters of data, especially when no other clusters existed. Then, once sections of data had been chosen, the class used their knowledge of fractions in order to pick ranges that encompassed enough of the entire distribution of data in order to be normal or usual. On April 6th, the class agreed that a third was sufficient.

It was hoped that during these class sessions, students would develop, critique, and revise data-based arguments about the nature of growth, variation, and differences between data sets about phenomena that they found personally intriguing. It was also a goal that the students would learn to see the data sets as distributions. Many other approaches to the analysis of the class sessions could have been taken, and maybe they

would have revealed clearer answers to these questions. For example, some other factors, such as time on topic, eye contact, or gestures could had been examined, and it is likely that entirely different results would have been found. Too, other questions could have driven the research. I did not come to the study with many preconceived ideas about what to look for, which was both a strength and weakness, allowing me to be open to what I would find but perhaps not providing enough structure to the study. Later efforts might expand and deepen these preliminary analyses to more fully inform the ways in which the students developed statistical understandings.

APPENDIX A

Order and Connectivity

138 coded turns/topics of talk
37 nodes
15 students and one teacher have at least one distinct speech turn
Coded students' turns range from 8 to 1

Coding Scheme follows. Bolded are the top eight nodes, by centrality. Centrality is defined by the sum of in and out connections, such as the number of distinct nodes connected to any given node.

Teacher talk and revoicing		
Code	Label	Description of Action
G1	Guiding question for comparing graphs	"Did the mystery ingredient make a difference?"
G2	Standards of discussion	"I want you to prove to me, to convince me, to show, test."
G3	Asks for clarification of conjecture/conceptualization	"Why would 29 be your choice?"
G4	Asks for clarification of technique/procedure	"Alex, when you said 28, were you thinking of 28 –and just the diameters on one side or the other of 28?" "So when he says the highest column, what do you mean?"
G7	Revoices conjecture/conceptualization	"So Alex's point is there are a lot of bubbles around 28 for the mystery solution. He says maybe that shows that the mystery solution made for better bubbles."
G8	Revoices technique/procedure	"So lots of times we looked at two numbers on either side of the number you were looking at." "Hillary's saying the regular blue solution has a lower range, from 5 to 37, and the mystery blue solution goes from 10 to 42."
G9	Asks for students to think a different way about conjecture/conceptualization or technique/procedure	"Who was thinking about the data a different way?"
G10	Asks students to make sense of differing conjecture/conceptualization	"What would you say to someone who listened to Hillary first and then heard Larissa say, the range is the same, but one starts at a higher point."

G11	Asks if ideas are sufficient to answer guiding question according to standards of discussion or asks about meaning	"Can we say for sure that the mystery solution worked better than the regular solution?" "So what would that make you want to say to me?" "Is that better performance?" "What does that tell you?"
G12	Asks students to think like one another in conjecture/conceptualization or technique/procedure	"What might Ross be thinking. Can someone else try to think like him?" "Did anyone else come to that in their thinking?"
G13	Asks students to make sense of differing technique/procedure	"CC: Anyone who isn't comfortable with that statement? Ross, did you get to that idea differently than Allison?"

Discussion topics that are considered to help answer a question about the data.		
Code	Label	Description of Action
C6	Thinks comparing whole range proves something	Prior comment: "CC: Hillary's saying the regular blue solution has a lower range, from 5 to 37, and the mystery blue solution goes from 10 to 42." Coded comment: "Lucas: Makes me think the blue solution makes smaller bubbles."
C8	Thinks comparing mode proves something	Prior comment: "CC: So the most bubbles that we blew that day that were one measurement. The measurement that came up the most." Coded comment: "Ryan: It means that the mystery ingredient gives bigger bubbles."
C10	Thinks comparing mode by proportion proves something	Prior comment: "CC: Does that mean anything? That a ninth is smaller than a sixth? What does that mean?" Coded comment: "Ross: That tells you that the mystery ingredient you can get larger bubbles than with the regular blue solution?"
C12	Thinks comparing same number on each proves something	"CC: Marissa is saying she looked at how many bubbles were 30 cm with the regular solution. There are 5 of them. Kind of low. I see 5 bubbles that measure 30 cm. We said there were 11 bubbles that measured 30 with the mystery solution. That might lead her to believe

		the mystery ingredient does make for bigger bubbles."
C13	Thinks comparing proportions/fractions past cutoff point on each graph proves something	Prior comment: "CC: Right. Here we got a third of the bubbles to be 30 or more. Is that better performance?" Coded comment: "Kids: Yeah!"
C14 (for now)	Creates alternate hypothesis	"Marissa: Maybe since they have the same blue solution, maybe just the corn syrup made it last longer instead of making bigger bubbles."

Procedures related to comparing frequency graphs		
Code	Label	Description of Action
P0	Short answer response to teacher question about technique.	Greg: "We had four"
P1	Noting end of range for one graph	"Greg: I noticed that in the mystery ingredient solution, we didn't have any fives or lower numbers in that. We had tens instead of fives."
P2 (teacher first)	Comparing ends of range (cut-off comparison? Slice comparison?)	Teacher: "So we didn't blow any bubbles with a diameter less than 10? How many less than 10 did we have with the regular solution?"
P3	Noting one graph's cluster, where a lot of numbers are, what we'd expect	"Ross: We have a lot in the column of 30. The 29 column isn't as high as the 30. The 28 has five in it. So then the one next to 28 has 5. If you put those together, you'd have a little higher number."
P4	Comparing clusters between graphs	"Alex: Well, I see that most of the numbers are around 28 instead of like 17."
P5	Noting whole range	Not used
P6 (teacher first)	Comparing whole range	"Hillary: The original solution had a lower range. From 5 to 37 and from 10 to 42."
P7	Noting mode descriptively	"Ryan: Like, on the most bubbles…" Only instance. Kill?
P8	Comparing mode descriptively	"Ryan: I was thinking about the blue solution, how the highest column is 20. With the mystery ingredient, the highest column was 30."
P9 (teacher first)	Comparing mode by count	"How many bubbles did we blow that were 30? Eleven. How many all had a measurement of 20?"
P10 (teacher first)	Comparing mode by proportion	"Eleven out of 63 measured 30 cm when we added the mystery ingredient."

P11	Creating proportions/fractions	"Ross: A third of 63 is 21. If you break the 21 into 3 groups, it's seven." "Matt M: I think that's a third of the bubbles."
P12	Comparing same number on each graph	"Marissa: If you look at the thirty with the regular blue solution, it's pretty low. There are many (more thirties) with the mystery ingredient."
P13 (teacher first)	Comparing proportions/fractions past cutoff point on each graph	"CC: So there we found 10 that were 30 cm or more, a sixth of all the bubbles. Here we just found 22 bubbles that were 30 or more cm in diameter."
P14 (teacher first)	Comparing count past cutoff point on each graph	"CC: Right. Here we got a third of the bubbles to be 30 or more. Is that better performance?"
P15	Noting median	"Alex: I think the halfway point would be about 27."
P16	Comparing median	"Alex: I was counting (the mystery solution) the bubbles up to 32. Remember, we broke them in half there."

Prior history, including how notations developed and are used		
Code	Label	Description of Action
Q1	General Context, such as how graphs were produced	"CC: So we added corn syrup and blew the same number of bubbles."
Q11	Reference to historical proportion data	"CC: On the 20th of Jan you guys figured out that about a sixth of the bubbles were 30 or more cm for the regular solution."
Q15	Reference to history of noting median	"CC: Alex wants you to remember when we looked at the regular blue and the pink, and we tried to find the middle of our data. So we could say half of our bubbles went this way and half went that way. So the number 32, that's where that came from."

Centrality of each node

	In -Connections	Out - Connections	Total In and Out Connections	div by 2 * # nodes
P11	9	8	17	0.23
G8	8	9	17	0.23
G11	6	9	15	0.20
P13	5	6	11	0.15
P0	5	6	11	0.15
G2	3	6	9	0.12
P2	4	4	8	0.11
G4	4	4	8	0.11
Q1	4	3	7	0.09
C13	4	3	7	0.09
P1	4	3	7	0.09
P10	4	3	7	0.09
G7	4	3	7	0.09
P12	3	3	6	0.08
G14	3	3	6	0.08
Q11	2	2	4	0.05
C14	2	2	4	0.05
P12	2	2	4	0.05
P14	2	2	4	0.05
P15	2	2	4	0.05
G1	2	2	4	0.05
C8	1	2	3	0.04
G12	2	1	3	0.04
Q15	1	1	2	0.03
C6	1	1	2	0.03
C10	1	1	2	0.03
C12	1	1	2	0.03
P4	1	1	2	0.03
P7	1	1	2	0.03
P8	1	1	2	0.03
P9	1	1	2	0.03
P16	1	1	2	0.03
G3	1	1	2	0.03
G9	1	1	2	0.03
G10	1	1	2	0.03
G13	1	1	2	0.03
P5	0	0	0	0.00

sum of the distinct nodes with connections to and from that particular node

APPENDIX B

Frequency of Mention.

Teacher Revoicing		
Code	Label	Description of Action
R1	Goal structure/ Goal posting	"Did the mystery ingredient make a difference?" "I want you to prove to me, to convince me, to show, test." "Can we say for sure that the mystery solution worked better than the regular solution?" "So what would that make you want to say to me?" "Is that better performance?" "What does that tell you?"
R2	Context	"CC: So we added corn syrup and blew the same number of bubbles." "CC: On the 20th of Jan you guys figured out that about a sixth of the bubbles were 30 or more cm for the regular solution." "CC: Alex wants you to remember when we looked at the regular blue and the pink, and we tried to find the middle of our data. So we could say half of our bubbles went this way and half went that way. So the number 32, that's where that came from."
R3	Assistance	"Who was thinking about the data a different way?" "What might Ross be thinking? Can someone else try to think like him?" "Did anyone else come to that in their thinking?" "CC: Anyone who isn't comfortable with that statement? Ross, did you get to that idea differently than Allison?"

Code	Label
N1	Noting highest or lowest case values
N2 (teacher first)	Comparing highest or lowest case values
N3	Noting whole range
N4	Comparing whole range
N5	Locate one graph's cluster
N6A	Comparing clusters between graphs descriptively
N6B	Comparing clusters between graphs by count
N6C	Comparing clusters between graphs by proportion/fraction
N6D	Comparing clusters between graphs by shape
N7	Noting mode (location of most of the high or low values in one distribution or other)

N8A	Comparing mode descriptively
N8B (teacher first)	Comparing mode by count
N8C (teacher first)	Comparing mode by proportion/fraction
N9	Locate a value from one distribution within the other
N10A	Comparing value descriptively
N10B (teacher first)	Comparing value by count
N10C	Comparing value by proportion/fraction
N11	Noting cut-off points within one distribution
N12A	Comparing between cut-off points between graphs descriptively
N12B (teacher first)	Comparing between cut-off points between graphs by count
N12C	Comparing between cut-off points between graphs by proportion/fraction
N12D	Comparing between cut-off points between graphs by shape
N13	Noting median
N14A	Comparing medians descriptively
N14B (teacher first)	Comparing medians by count
N14C	Comparing medians by proportion/fraction
N14D	Comparing medians by shape
N15	Discuss proportion or fraction of separation or overlap of the two distributions
N16	Creates alternate hypothesis
N17 (teacher first)	Expectation of typical number
N18	Noting spaces, gaps
N19	Creating/preparing proportions/fractions

REFERENCES

Erickson, F. and Schultz, J. (1977/1997). When is a context? Some issues and methods in the analysis of social competence. In M. Cole, Y. Engestrom, & O. Vasquez (Eds.), Mind, culture, and activity: Seminal papers from the Laboratory of Comparative Human Cognition (pp. 22-31). Cambridge, UK: Cambridge University Press. Originally appeared in *Quarterly Newsletter of the Laboratory of Comparative Human Cognition 1*, 5-10, 1977.

Hall, R. and Rubin, A. (1998). …There's five little notches in here: Dilemmas in teaching and learning the conventional structure of rate. In J. Greeno & S. Goldman (Eds.) *Thinking Practices in Mathematics and Science Learning*. Lawrence Erlbaum Associates: Mahway, New Jersey. pp. 189-236.

Konold, C. and Khalil, K. (2003). If U Can Graff These Numbers – 2, 15, 6 – Your Stat Literit. Paper presented a the Annual Meeting of the American Educational Research Association.

Lehrer, R., & Schauble, L., Eds. (2001). *Investigating real data in the classroom: Expanding children's understanding of math and science* . New York: Teachers College Press.

Lehrer, R., Strom, D., & Confrey, J. (2002). Grounding metaphors and inscriptional resonance: Children's emerging understanding of mathematical similarity. *Cognition and Instruction, 20* , 359-398

O'Connor, M.C., Michaels, S. (1996). Shifting participant frameworks: orchestrating thinking practices in group discussion. In D.Hicks (Ed.), *Child Discourse and Social Learning.* Cambridge: Cambridge University Press. pp. 63-102.

Orsolini, M. and Pontecorvo, C. (1992) Children's talk in classroom discussions. *Cognition and Instruction 9*(2), 113-136.

Strom, D., Kemeny, V., Lehrer, R., & Forman, E. (2001). Visualizing the emergent structure of children's mathematical argument. *Cognitive Science, 25* , 733-773.