Chapter 15

SECONDARY TEACHERS' STATISTICAL REASONING IN COMPARING TWO GROUPS[1]

Katie Makar[1] and Jere Confrey[2]
University of Texas at Austin, USA[1], and Washington University in St. Louis, USA[2]

OVERVIEW

The importance of distributions in understanding statistics has been well articulated in this book by other researchers (for example, Bakker & Gravemeijer, Chapter 7; Ben-Zvi, Chapter 6). The task of *comparing* two distributions provides further insight into this area of research, in particular that of variation, as well as to motivate other aspects of statistical reasoning. The research study described here was conducted at the end of a 6-month professional development sequence designed to assist secondary teachers in making sense of their students' results on a state-mandated academic test. In the United States, schools are currently under tremendous pressure to increase student test scores on state-developed academic tests.

This paper focuses on the statistical reasoning of four secondary teachers during interviews conducted at the end of the professional development sequence. The teachers conducted investigations using the software Fathom™ in addressing the research question: "How do you decide whether two groups are different?" Qualitative analysis examines the responses during these interviews, in which the teachers were asked to describe the relative performance of two groups of students in a school on their statewide mathematics test. Pre- and posttest quantitative analysis of statistical content knowledge provides triangulation (Stake, 1994), giving further insight into the teachers' understanding.

[1] This research was funded by the National Science Foundation (NSF) under ESR-9816023. The opinions expressed in this chapter do not necessarily reflect the views of NSF.

D. Ben-Zvi and J. Garfield (eds.),
The Challenge of Developing Statistical Literacy, Reasoning and Thinking, 353–373.

WHY STUDY TEACHERS' REASONING ABOUT COMPARING TWO GROUPS?

Statistics and data analysis are becoming increasingly important in our society for a literate citizenry. As such, many schools have begun to incorporate statistics and data analysis into their curriculum, beginning as early as Kindergarten (TERC, 1998). Although many schools are increasing their emphasis on statistics, very few are taking sufficient steps to help teachers master the statistics they are expected to teach. Professional development typically provided to teachers by their schools gives mathematics teachers little opportunity to improve their statistical content knowledge beyond evaluation of central tendency and simple interpretation of graphs and tables, while university statistics courses are rarely aimed at content teachers feel is relevant. Furthermore, U.S. teachers have little experience with data analysis and inferential statistics, yet in a time when teachers are under increasing pressure to improve student scores on state-mandated tests, teachers are required to make instructional decisions based on large quantities of data about their students' performance. Given that teachers are both the target and the vehicle of reform (Cohen & Ball, 1990), it is vital that we consider teachers' facility in statistical reasoning as well as possible vehicles for helping teachers improve their conceptual understanding of the statistics they are expected to teach. Enhanced understanding of teachers' statistical reasoning will help professional development leaders better design conceptual trajectories for advancing teacher reasoning in statistics, which should ultimately improve student understanding in probability and statistics.

Investigations involving comparing groups provide a motivational vehicle to learn statistics (see, for example, Konold & Pollatsek, 2002): They are steeped in context, necessitate a focus on both central tendency and distribution (for various aspects of distributions, see Chapter 7 this volume), and provide momentum for the conceptual development of hypothesis testing. Furthermore, tasks involving group comparisons are rich enough to be accessible to a broad array of learners at varying ages and levels of statistical understanding. Comparing distributions can be an interesting arena for researchers to gain insight into teachers' statistical reasoning, and in particular gave us an opportunity to understand teachers' reasoning about variation in a more sophisticated context.

Several curriculum projects make use of group comparisons as an avenue to teach statistical reasoning. At the elementary level, comparing two groups can be used to introduce the concepts of data and graphing, providing students with important early experiences in viewing and reasoning with distributions. For example, a first-grade curriculum (TERC, 1998) introduces primary students to distributions by having them compare and qualitatively describe the distribution of their classmates' ages to that of their classmates' siblings. Middle school students are able to build on earlier experiences with data and start to focus on descriptions of distributions: measures of center and spread, shapes of distributions, as well as gaps and outliers. For example, a sixth-grade curriculum puts these skills into a meaningful context for students by having students compare "typical" heights of males and females in their class, examining measures of center, describing the shapes of two distributions, and looking at gaps and outliers (Lappan, Fey, Fitzgerald, Friel, & Phillips, 1998). For older students, more open-ended designs and conventional descriptions of statistical variation can be introduced, which will help students build a foundation for inferential statistics or to inform debate of issues in light of available data.

At a wide variety of grade levels and settings, comparing groups has the potential for giving students authentic contexts to use data to answer meaningful questions, thus motivating the power of data in decision making. However, in order for teachers to provide these kinds of tasks for their students, they need to develop their own statistical understanding. Heaton and Mickelson (Chapter 14, this volume) described the experience of an elementary teacher's struggle to develop her own statistical reasoning as she worked to merge statistical investigations into the existing school curriculum. This chapter will examine statistical reasoning in secondary teachers as they build their statistical content knowledge through investigations of student assessment data, in particular the role of variation in considering what it means to compare two groups. (For additional discussions of the teachers' reasoning with data, see Confrey & Makar, 2002; Makar & Confrey, 2002.)

PREVIOUS RESEARCH ON COMPARING TWO GROUPS

Within the world of statistics, much concern is placed on making comparisons, either direct or implied. Whether the difference is between brands of peanut butter, or housing prices compared to last year, comparisons form the very fabric of research and of *principled* arguments (Abelson, 1995):

> The idea of *comparison* is crucial. To make a point that is at all meaningful, statistical presentations must refer to differences between observation and expectation, or differences among observations. Observed differences lead to why questions, which in turn trigger a search for explanatory factors ... When we expect a difference and don't find any, we may ask, "Why is there *not* a difference?" (p. 3)

Lehrer and Schauble (2000), in their work with children's graphical construction, indicate that young students "are often disconcerted when they find a discrepancy between the expected value of a measure and its observed value" (p. 104). Watson and Moritz (1999) argue that comparisons of data sets provide a meaningful backdrop for students to gain a deeper understanding of the arithmetic mean as well as strong intuitive approaches to compare groups through balancing and visual strategies, "hopefully avoiding the tendency to 'apply a formula' without first obtaining an intuitive feeling for the data sets involved" (p. 166).

The task of comparing groups appears in the literature as an impetus for students to begin to consider data as a distribution instead of focusing on individuals, in addition to motivate students to take into account measures of variation as well as center (Konold & Higgins, 2002). Lehrer and Schauble (2000) found that as older students solved problems in which they compared two distributions, they began to look at both centrality and dispersion. In their study, comparing groups served as an impetus for students to gain an appreciation for measures beyond center. For example, they report on a group of fifth graders who, when experimenting with different diets for hornworms, found that the hornworms in the two treatment groups showed differences not only in their typical lengths but also in the *variability* of their lengths. This caused the students to speculate and discuss reasons why the lengths in one group varied more, showing that "considerations of variability inspired the generation of explanations that linked observed patterns to mechanisms that might account for them" (p. 129).

Examining the context of a problem is critical for understanding group comparisons. Confrey & Makar (2002) discuss the role of context in statistical learning while examining the process of teachers' inquiry into data. In one activity they describe, teachers examined several pairs of graphs void of context and reasoned about comparisons between graphs in each pair at a very superficial level in a discussion that lasted only about 5 minutes. However, when the same graphs were examined again in light of a context relevant to the teachers (quiz scores), a much more in-depth analysis took place in a discussion lasting 40 minutes. This discussion was the first time in their study that the teachers articulated variation in a distribution as being useful. When the teachers could compare distributions in a personally meaningful context, they began to gain a more robust understanding of distribution. Similarly, Cobb (1999) found that by comparing the distributions in the context of judging the relative lifespan of two types of batteries, students were compelled to consider what it meant for one battery to be preferred over another—does one consider overall performance, or consistency? Here, students negotiated a purposeful reason to consider variation in the context of what constitutes a "better" battery.

Comparing two groups also becomes a powerful tool in light of its use toward a consideration of statistical inference. Watson & Moritz (1999) argue specifically that comparing two groups provides the groundwork "to the more sophisticated comparing of data sets which takes place when t-tests and ANOVAs are introduced later" (p. 166). Without first building an intuitive foundation, inferential reasoning can become recipe-like, encouraging black-and-white deterministic rather than probabilistic reasoning. "The accept-reject dichotomy has a seductive appeal in the context of making categorical statements" (p. 38, Abelson, 1995). Although formal methods of inference are not usually a topic in school-level statistics content, an ability to look "beyond the data" (Friel, Curcio, & Bright, 2001) is a desired skill. Basic conceptual development of statistical inference can lead to assistance in understanding one of the most difficult, but foundational concepts in university-level statistics: sampling distributions (delMas, Garfield, & Chance, 1999).

RESEARCH DESIGN AND METHODOLOGY

The research described in this chapter was part of an NSF-funded research project developed and carried out by a research team at the Systemic Research Collaborative for Education in Math, Science, and Technology at the University of Texas at Austin. Although this chapter focuses on the results of interviews taken at the end of the study, the experience of the participants in the research project is key to understanding their background knowledge and experience in statistical reasoning. It should be noted that research on these teachers' statistical reasoning was not the purpose of the workshop, which was to examine the effects of the professional development sequence within a larger systemic reform project (Confrey, in preparation). The authors saw an opportunity, after the workshops were planned, to examine the teachers' statistical reasoning through a set of clinical interviews. This chapter is the result.

The 6-month professional development research project took place in two phases: 18 contact hours of full-day and after-school meetings, followed by a 2-week summer institute. The project was conceived as a mathematical parallel of the National Writing Project, where teachers focus on their own writing rather than how to teach writing. A mission of the National Writing Project (2002), and a belief that was fundamental to our study, is that if teachers are given the opportunity to focus on their own learning of the content that they teach—to *do* writing, or mathematics, in an authentic context—they will better understand the learning process and hence teach with greater sensitivity to students' conceptual development (Lieberman & Wood, 2003). Our professional development sequence was designed under the assumption that if mathematics teachers are immersed in content beyond the level that they teach, and developed through their own investigations as statisticians within a context that they find compelling and useful, then they will teach statistics more authentically and their increased content knowledge will translate into improved practice.

During the professional development sequence, teachers learned a core of statistical content: descriptive statistics and graphing, correlation and regression, sampling distributions, the Central Limit Theorem, confidence intervals, and basic concepts of statistical inference. These concepts were not developed formally, as they would be in a university course; rather, teachers were given extensive experience with sampling distributions through simulations in order to (a) help them understand concepts of sampling variation that we thought was critical to their working with data and (b) give them access to powerful statistical ideas. Statistical concepts were introduced only as they were needed to make sense of the data; many of the teachers already had at least a working knowledge of descriptive statistics and graphing, as indicated by their statistics pretest.

During the workshops and summer institute, teachers conducted increasingly independent investigations focused on the analysis of their students' high-stakes state assessment data. For the teachers, this was a compelling context in which to learn statistics. In Texas, there is much emphasis on the Texas Assessment of Academic Skills (TAAS, www.tea.state.tx.us), the high-stakes state assessment where students and schools are held accountable for their performance on the battery of tests. Teachers felt they would be empowered if they were able to interpret TAAS data instead of having to rely on experts to tell them what the data meant and what actions the school needed to take in order to raise test scores. Because many of the "lessons" we wanted them to gain from working with data involved sampling variation, we felt it critical to give them enough experience to develop an intuition about this type of variation.

Many of the investigations were supported by the use of the statistical learning-based software, Fathom (Finzer, 2000), to examine data directly as well as to create simulations to test conjectures. The software allowed teachers to fluidly and informally investigate relationships in the data because of the ease with which Fathom creates graphs through a drag-and-drop process. Most statistical software tends to be like a "black box" with a purpose that supports a data-in, results-out mind-set that can work to encourage misconceptions in early learners who expect to find definitive answers in the data. Fathom insists that users build simulations in the same way that one would construct a sampling distribution: by creating a randomization process, defining and collecting measures from each randomization, and then iteratively collecting these measures. The levels of abstraction that make sampling distributions and hypothesis testing so difficult for university students (delMas et al., 1999) are not avoided in Fathom, but made transparent through creating visual structures that parallel these processes, allowing users to better visualize the concepts underlying the abstract nature of a sampling distribution. Fathom was also a powerful tool for analysis and supported the use of authentic data, even reading data directly from websites, thus empowering teachers to greater access to the many data sets that are available on the Internet.

During the workshops with the teachers, we often used sampling distributions to illustrate and investigate statistical concepts—properties of the normal distribution, the Central Limit Theorem, the effect of sample size on sampling variability, the null hypothesis, p-values, and hypothesis testing. In addition, these statistical concepts were applied during investigations of relationships in the data. It is important to note that we did not focus explicitly on group comparisons during the professional development workshops; we did not formally develop a list of procedures for comparing groups, nor had the teachers seen a task similar to the one we asked them to perform for the interview. During the workshops, the teachers did engage in two structured activities in which comparing two groups was central. The first activity took place in the early stages of the professional development program, when teachers were first learning the software and investigating descriptive statistics. In this activity (Erickson, 2001, p. 206), a sample of student scores on the Scholastic Aptitude Test (SAT; a national test many U.S. students are required to take as part of their college application) and the grade point averages of males and females were examined and informally compared. The second activity, *Orbital Express* (Erickson, 2001, p. 276), took place in the final week of the summer institute when investigating the concept of a null hypothesis and working with more advanced features of the software. In this activity, teachers dropped two types of wadded paper and attempted to hit a target 10 feet below. The distance each wad fell from the target was then entered into one column in Fathom, and the type of paper thrown was entered in a second column. Using the *scramble attribute* feature of the software, the values in the second column (type of paper) were randomized, simulating a null hypothesis, and the difference in the median of each group was calculated. This process was repeated 100 times to create a "null hypothesis" distribution, showing the amount of variation in the differences between the medians of the groups that might be expected just by chance. The difference found in the original sample was then examined in light of this "null hypothesis" distribution.

SUBJECTS AND DATA COLLECTION

This chapter focuses primarily on four secondary mathematics teachers from Texas who took part in the professional development program just described. Two of these participants joined the project later and were part of an abbreviated repeat of the first phase of the professional development sequence. One of the four subjects was a preservice teacher while the other three were experienced, credentialed secondary mathematics teachers who taught 13- to 16-year-old students. Two of the teachers had obtained a university degree in mathematics, the preservice teacher was working on her mathematics degree, and the remaining teacher had a degree in the social sciences. The two men and two women consisted of one Hispanic and three non-Hispanic whites. All but the preservice teacher had taken a traditional introductory statistics course 5–15 years previously during their university coursework. These were the only four teachers who took part in Phase II of the project (the summer institute), due partly to a change in the administration at the school and scheduling conflicts with teachers teaching summer school.

Data collected on the subjects included a pre-post test of statistical content knowledge, which was analyzed using a t-test in a repeated measures design. In addition, all of the sessions with the teachers were videotaped. Interviews were conducted at the end of the study in which participants were asked to compare the performances of two groups, and which will comprise the main source of data for this chapter. The interviews were videotaped, and major portions were transcribed and then analyzed using the qualitative methodology of grounded theory (Strauss & Corbin, 1998). Under this methodology, the transcripts were first subjected to open coding in the software NVivo (QSR, 1999) to capture the phenomenon observed in the teachers' own words and actions and to allow potential categories to emerge from the data that would describe strategies, mind-set, and other insights into how the teachers were comparing groups. Secondly, initial categories were organized into hierarchical trees and subjected to axial coding to begin to tie common themes into larger categories. Finally, the data were analyzed with selective coding to further investigate various dimensions of the categories and better describe the phenomenon observed.

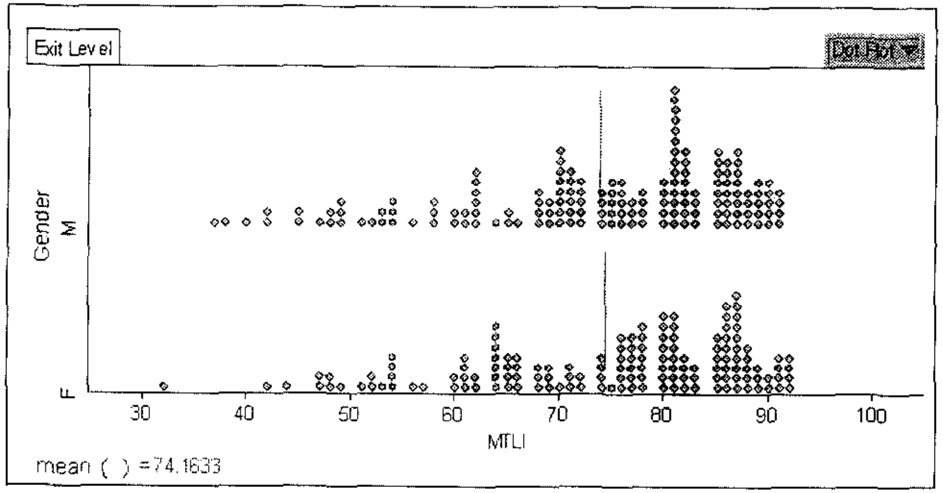


Figure 1. A dot plot of student test scores with means plotted in Fathom created by one teacher.

INTERVIEW TASK

In the interviews, which took place during the last two days of the summer institute, subjects were given a raw data set of student TAAS scores from a hypothetical low-performing high school and asked to use Fathom to compare the performance of males and females in the school. Although all of the data used in the example was not from a single school, it was in fact authentic student data created from a compilation of scores drawn from several schools in Texas. Figure 1 shows a graph similar to ones that each of the teachers initially created in Fathom from the given data. The MTLI on the horizontal axis of this graph is the Mathematics "Texas Learning Index" on TAAS; a MTLI score of 70 is considered passing. In the context in which the state high-stakes test exists, it is not just the means that are relevant to consider. Schools are held accountable for the proportion of students who pass the TAAS test, so it is also important to consider the proportion passing for each group. In a second task during the interview, subjects were asked to investigate the low-performing status of the school, based on analysis of the performance of the state-defined ethnic subgroups within the school and to make a campus-based program recommendation to the principal for the following year. The analysis in this chapter focuses on the first interview task.

RESULTS

In analyzing these teachers' reasoning about comparing two groups, we assumed that the professional development sequence had an impact on their content knowledge. Rather than examine teachers' reasoning about comparing two groups with teachers who had little experience with data or diverse backgrounds in statistical content knowledge, we chose to examine teachers who had developed their statistical understanding through rich experiences as investigators. We recognize that the reasoning ability of this group, therefore, is not *typical* of secondary teachers, but what *could* occur after a relatively short period of professional focus on building conceptual understanding and experience with powerful statistical ideas. The overarching purpose of the professional development was to give them rich experiences as investigators with school data. We did discuss many concepts in inferential statistics, but the majority of these more advanced concepts (e.g., t-tests, confidence intervals, null hypothesis, *p*-values) were experienced through simulations on a conceptual, not formal level.

To measure whether the content that was taught had an impact on teachers' understanding and to assess the level of statistical content knowledge at the time of the interviews, a pre-post test of content knowledge was given to teachers. The result of the analysis is given in Table 1. The data summary shows significant growth ($\alpha = 0.05$) in their overall content knowledge as well as for two individual areas (Sampling distributions and Inference), even though the number of teachers in the study was small ($n = 4$).

Table 1. Results of pre-post test of statistical content knowledge using a t-test and repeated measures design, *n*=4

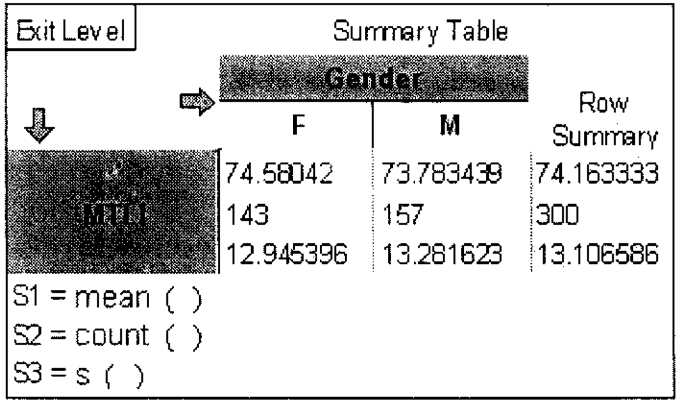
Topic	Pretest Mean Percent Correct	Posttest Mean Percent Correct	Difference	t	p-value
Descriptive Statistics	61%	79%	18%	2.0	0.14
Graphical Representation	75%	83%	8%	0.5	0.63
Sampling Distribution	8%	75%	67%	4.9	< 0.01
Inference and Hypothesis Testing	6%	59%	53%	18.0	< 0.01
Overall	**35%**	**71%**	**36%**	**6.8**	**< 0.01**

While quantitative methods could be used to measure content knowledge, it was necessary to use qualitative methods to better understand teachers' *statistical reasoning* about comparing two groups. During the qualitative analysis, 20 initial categories were organized into four final categories—conjectures, context, variation, and conclusions—by collapsing and generalizing other categories. Finally, the researchers developed a preliminary framework for examining statistical reasoning (Makar & Confrey, 2002). This chapter focuses on elements that are specific to teachers' reasoning about comparing two distributions. Of special interest are teachers' conceptions of variation, which bring with them several issues that are unique to the task of comparing groups. In examining these teachers' descriptions of the two distributions, it will be interesting to note how they choose to compare the characteristics of the each distribution. For example, do they see these measures as absolute, or do they recognize the possibility of inherent error? That is, do they view any small differences in these measures quantitatively as absolute differences, or do they indicate a tolerance for variation in these measures, so that if these students were tested again, they might expect to see slightly different results?

EXAMPLES OF TEACHERS' REASONING ABOUT COMPARING DISTRIBUTIONS

In this section, we discuss the data from interviews with the four subjects from the second phase of the study: Larry, Leesa, Natalie, and Toby. These four teachers were the only four participating in Phase II of the study, the 2-week summer institute.

The first transcript we examine is Larry's, who has taught middle school math for 6 years. He has an undergraduate major in mathematics and is certified to teach mathematics at the middle and high school levels. Larry's initial portrayal of his comparison of the two distributions began with a visual evaluation of the similarity of their dispersion, then a numerical description of the means and standard deviation of each of the two distributions. He finished with a comparison of these measures:



Exit Level

Summary Table

Gender

	F	M	Row Summary
MTH	74.58042	73.783439	74.163333
	143	157	300
	12.945396	13.281623	13.106586

S1 = mean ( )
S2 = count ( )
S3 = s ( )

Figure 2. Larry's summary table in Fathom. The three rows correspond to the values of the mean, count, and standard deviation for the females, males, and then total group.

Larry: I'm just first dropping them, both of them in a graph (Figure 1), the math scores of the males and females. Um, both of them seem to be fairly equally distributed, maybe. I'm going to try and find the means of each one (mumbles). I'll just graph them, then. Hmm. So they're fairly close … I'm pulling down a summary table (Figure 2) so I can actually find a number for each one. The, uh, so I actually get a real number. So it's giving me the count and mean of each one. Also, here I can find out, uh (mumbles), I can find the standard deviation of each one to see how close they are from the mean.

KM: And how, how will that help you?

Larry: Well, if, even if I didn't see the graph, I can look at the females are even a little tighter, around a higher mean.

KM: OK.

Larry: On both sides. As opposed to the men, also—that are a little more spread around, around a lower average.

Larry later considered the difference of the means more directly, estimating the difference from the figure:

Larry: Even though they're going to be very close, I, I think, I, I mean, there's not a great difference between the men and the women. But the women look like they scored maybe one or two points higher.

Larry here acknowledged that the difference between the means was very close, but did not interpret the difference as anything other than a 1- or 2-point difference. At the end of the first part of the interview, Larry informally compared the extreme values of the two distributions, as well as their means and proportion passing, to summarize his analysis:

KM: Just describe for me, if you were going to compare those two groups, the performances of those two groups. Describe the similarities and differences.

Larry: OK. The females have a larger range, because the lowest score and the highest score are—the lowest score of the females is lower than the lowest score of the males, and the highest score of the females is higher than the highest score of the males. Uh, so the, the range is higher. Yet, still the, the mean score is higher than the average score of each of—, the females is higher than the average score of the males.

Larry's comparisons consisted of independent descriptions of each distribution along with direct comparisons of center and dispersion. While he considered the variability of each distribution, he did not indicate a sense of the variation between the measures of the two distributions—that is, he compared the means and dispersions of the two distributions qualitatively or in absolute terms. He concluded that the mean of the females was higher than that of the males by observing an estimated 1- or 2-point difference in the means. While he asserted that these were close, Larry indicated no particular inclination to investigate whether the difference in the means was significant.

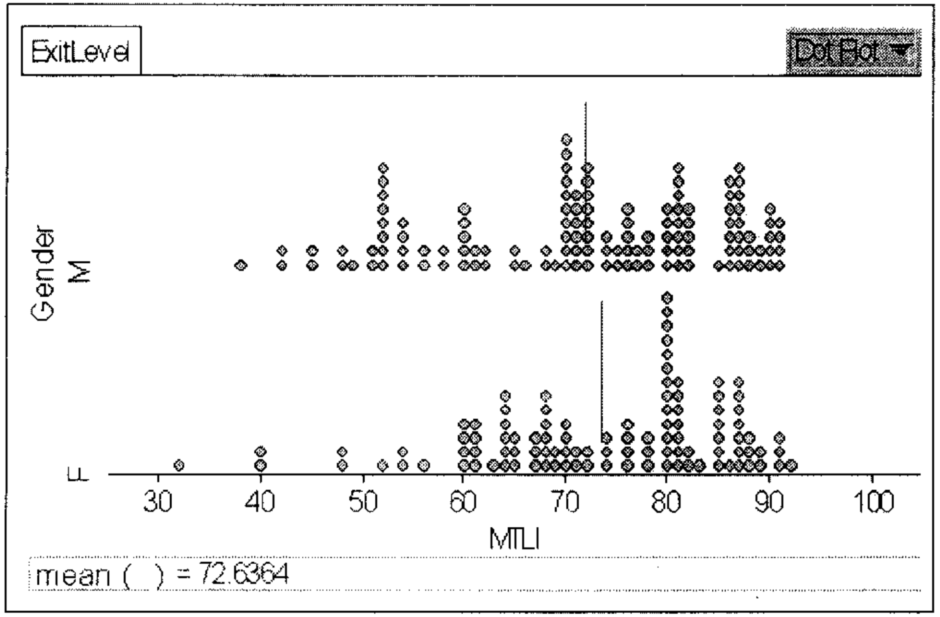


Figure 3. Leesa's initial dot plot of Gender vs. MTLI (TAAS math scores).

Leesa has taught middle school mathematics for 7 years, has undergraduate majors in the social sciences, and is the chair of her mathematics department. Her initial description included comparisons of the shape, range, maximums, and means of each distribution (Figure 3; note that Leesa's data set, and those of the other two teachers who follow, are slightly different than Larry's):

Leesa: OK, um, let's see. This looks skewed to the left [pointing to the top distribution]. Well, they both look skewed to the left. Uh, the range of the males looks like it goes from about nine[ty]—well, they're about the same. There's a bigger range in the female performance because of this, this one student right here who has got a 32.

KM: OK.

Leesa: Um. A high score is a 92 on female and on male it's a 91. Um, and then I can also, I can also go and find the mean. And then, [pause] the edit formula and plot the mean on both of those [Leesa selects the graph, chooses "edit formula" from the context menu, and plots the means on the graph in Fathom]. So for the females it looks like their line is about 72.6, no, 73 [Leesa moves the cursor close to the mean line for the females and reads the location of the cursor in the corner of the Fathom screen]. And then for the males, it looks like about 72.

KM: OK.

Leesa: So the average female score is just a little bit higher than, than the average male.

Leesa seemed to view the measures she stated as descriptions of each distribution separately, although she made some comparison in these measures, indicating some tolerance for variation in her qualitative description of the range of the distributions as "about the same." She did not hold on to this view, however, when she moved from qualitatively to quantitatively comparing the distributions. For example, while she found that the mean for females was higher than for males, she did not indicate whether she interpreted this 1-point difference as their centers being about the same. In the interview, Leesa went on to compare the proportion of each group that passed [63% of the females passed compared to 68% of the males]. She noted that alternative measures were giving her seemingly conflicting information about the relative performance of the two groups, stating, "More boys passed than girls when you look at percentages, but, and the mean score of the girls is higher."

When asked to sum up whether she saw any difference between the performance of males and females at the school, Leesa considered using a sampling distribution this time to provide evidence that the difference between the two groups was not significant. However, her attempt to do so included a laundry list of methods we had used in the summer institute to examine variability, including a reference to "what we did this morning," which was a procedure in Fathom called *stacking* that probably would not have been useful in this situation:

KM: So can you, you say whether one performed better than the other?

Leesa: No.

KM: What evidence could you give me to, that there wasn't any difference, what would you say?

Leesa: Um, I can do that test hypothesis thing. Um, I could do one of those, um, like what we did this morning, the sample and see if there was any—How many students were there? 231?

KM: Uh-huh.

Leesa: I could do a smaller sample and just kind of test and see if, see what the means look like each time … OK, then when you do standard deviation—is that really going to help me here? Because, let's plot it and see what it looks like [Leesa plots a marker on her graph in Fathom, one standard deviation above the means].

KM: OK, why do you think that might give you something, or are you just going to see—

Leesa: Um. I just want to see if this, if this mean, if this mean [pointing to the females in Figure 3]—

KM: Uh-huh?

Leesa: —falls within one standard deviation of the top mean [the males].

KM: *Do you think it will?*

Leesa: Yes. (pause) So it's not like it's a huge difference, I guess.

KM: So what does checking where the standard deviation, what does that tell you? What does that measure? Try and think out loud.

Leesa: Um, OK. Standard deviation means that most of the scores are going to fall within there [the interval within one standard deviation of the mean]. So, I don't really see how that—OK, I understand what we were doing yesterday when we had the standard deviation and then, you know, when we had, uh, when we looked to see if that would, if that was really weird. And if it fell outside the standard deviations, when we looked at z-scores and they were really high, if it fell way out here, then we know that was something, not typical.

KM: OK.

Leesa: OK, but since this, these are so close together, and it falls within, you know, that that's pretty typical and, it might go either way.

Unlike Larry, Leesa indicated a tolerance for variation between the measures she used to compare the two groups. Even though the means of the groups *were* different, she acknowledged that the difference was not enough for her to decide whether one group performed better than the other. She struggled, however, with providing evidence that the difference was not meaningful. Her explanation contained a hybrid of concepts related to the distribution of scores and that of a sampling distribution, together with a list of procedures she might try.

Natalie, a preservice teacher and mathematics major with no previous statistical coursework, immediately took a less deterministic stance in her comparison of the performances of males and females on the TAAS test at the hypothetical school. Natalie initially created a dot plot of the data (similar to the one in Figure 3), then changed it to a histogram. She then created a summary table in Fathom to calculate the means and standard deviations of the MTLI score for each gender:

Natalie: It looks like the mean for the females is a couple of points higher than the mean for the males [pointing to the summary table], but whether or not that's significant, I don't know yet ... I don't think they're very different. It just happens to come up a little bit higher, but the standard deviation is 13 points, so 2-point difference isn't all that much ... The, the range looks about the same to me, I mean, there's a few extra down there in the females, but I don't think that's very significant. They look pretty similar ... I don't think they're, they're very different.

Natalie immediately considered whether the difference she was seeing in the means was significant and went on to conclude that the 2-point difference in the means of the two groups was probably not significant, relative to the distribution of scores. She compared the 2-point difference in means to the standard deviation rather than to their standard error, since she did not consider the size of the group in her interpretation of significance. It's possible that she was considering not *statistical significance,* but a more informal notion of a *meaningful difference* relative to the distribution of scores.

The final interview was with Toby, an experienced high school teacher who has been teaching for over 10 years. Toby's initial comparison between the two groups (creating a graph similar to Figure 3) was based on a visual interpretation, before considering a numerical comparison:

KM: Describe to me what you see, compare those two groups.

Toby: Well, just by looking at that I would say that the, the men scored better than the women. Um, then I would probably drop, um, means in there. Um, probably get an idea of what that was. Uh, 74, closer to 74, and that was 72. Not, not that much difference. They're about the same.

KM: The same?

Toby: Yes.

KM: And you're basing that on?

Toby: Uh, that the means are pretty close together and that, there's about, uh, there, there are no real outliers ... The females averaged higher, um, there's one kind of low one out there, but there's not that much, they're a pretty close group, pretty closely grouped. If we had to go farther, we might, now I don't know how big this set is but I used all of the data, so.

KM: So if somebody said, you know, is there any difference between these two groups?

Toby: Well, to get, well, we could do those things like what we've been doing. Uh. How, how many is this? Uh, only 230. Well, uh. And they're all there. We can do one of those things about, you know, pick 50 of them at a time, find that average, pick 50 at a time, find that average, pick 50 at a time, and then look at that, uh, the average of those.

KM: Uh-huh.

Toby: OK. And, uh, that's going to tend to squish the data together, and, towards whatever the real mean of that data is, but it would also give me a, uh, idea of, of the spread or the vari—how, how the highs and lows were.

KM: OK.

Toby: Of that spread.

Toby also interpreted the difference that he found in the means as being "about the same," indicating he, too, possessed an expectation of variation between the measures of the two groups. Toby also recognized that a sampling distribution of some kind would help support his assertion that the difference between the two groups was not significant, but he had similar difficulties determining how to set up a sampling distribution or how to incorporate the sizes of the groups.

DISCUSSION

In examining teachers' reasoning about comparing distributions, we found that teachers were generally comfortable working with and examining traditional descriptive statistical measures as a means of informal comparison. An interesting contrast occurs, however, when we consider teachers' conceptions of *variability* when reasoning about comparing two distributions. As indicated in the literature, variability is an under-researched area of statistical thinking (Meletiou, 2000). Yet attitude toward variability could provide an important indication of statistical mindset (Wild & Pfannkuch, 1999). Having an understanding and tolerance of variability encompasses a broad range of ideas. In examining the concept of variability with only one distribution, one considers the variation of values *within* that distribution. However, descriptive statistics for a single distribution are often viewed without regard to variability of the statistical measures themselves. With one distribution, there is little motivation to consider or investigate possible sources of variation in the measures drawn. Comparing distributions creates a situation where one is pushed to consider measures less deterministically. Depending on the measure that dominates the comparison (often a mean), how does one interpret differences found in measures *between* groups? That is, how does one determine whether any difference between the dominant measures is meaningful or significant? Further, how do teachers manage the distinction between these two kinds of variation? By considering variation between distributions, we are encouraged to consider sources of variation in these measures. In this chapter, we discuss three different ways that teachers considered issues of variability when reasoning about comparing two distributions: (1) how teachers interpreted variation *within a group*—the variability of data; (2) how teachers interpreted variation *between groups*—the variability of measures; and (3) how teachers *distinguished* between these two types of variation.

In the interviews, all four teachers knew that scores within each distribution of scores would possess variability—that is, they did not expect the data in the distribution of scores would all have the same value. Teachers' conceptions of this *within-group variation* were heard in their descriptions of shape, distribution, outliers, standard deviation, range, "domain" (maximum and minimum values), and "whiskers" on a box plot (not included in the preceding excerpts, but used by two of the teachers). Additional qualitative descriptions included statements about a distribution being "tighter" or "more spread out." Commonly, teachers calculated the standard deviation of each set almost immediately and somewhat automatically.

While all of the teachers clearly recognized variation *within* a single distribution, they articulated a variety of meanings about variation *between* two distributions. From our interaction with them in the workshops, we anticipated they would demonstrate their view of between-group variation by acting in one of four ways: (a) by calculating descriptive statistics for each group without making any comparisons; (b) by comparing descriptive statistics (e.g., indicating a difference in magnitude or that one was greater then the other); (c) by first comparing the descriptive measures of the two distributions as described earlier, then indicating whether they considered the difference to be meaningful by relying on informal techniques or intuition; or (d) by investigating whether the differences they found in the measures to be statistically significant using a formal test, such as the randomization test the teachers carried out during the *Orbital Express* activity (Erickson, 2001, p. 276) using the *scramble attribute* feature in Fathom, which randomizes one attribute of the data.

In addition to describing the variation *within* each distribution separately, the teachers typically reported some aspect of the similarity or differences in the measure of dispersion between the two distributions, by comparing range or standard deviation. They may also have compared shapes or means, for example, by noting that the mean of the females' scores was 2 points higher than that of the males. In some cases, teachers indicated an intuition about variation between measures, but struggled to quantify the evidence for their observations. One reason for our perception that teachers had difficulty in quantifying variation between distributions may be that the participants felt they were being pushed to provide evidence of what seemed to them to be an obvious example of two distributions that were "about the same." Perhaps to the teachers, the sameness could be seen visually, and they would not feel compelled to provide evidence of this observation under less test-like circumstances.

Two of the teachers, Leesa and Natalie, attempted to formally test whether the difference in the means of the two distributions was significant using some form of a standard deviation taken from the data distributions. Furthermore, Toby, as well as Leesa, checked the size of the population to see if it was "large enough" to draw samples from, perhaps recalling that several times during the workshop they had created sampling distributions by drawing random samples from a state data set of 10,000 student test scores. Neither of them, however, used the size of the data set in determining whether the difference in means between the males and females was significant. Overall, the three who considered using a sampling distribution struggled to understand the circumstances under which using one would be helpful nor were they able to separate the variability in the distributions of the data sets from that of the related sampling distribution, confirming that this is a very difficult concept to understand in statistics, consistent with the findings of delMas, Garfield, and Chance (1999).

Using Confrey's (1991, 1998) concept of *voice and perspective*, the authors brought to the research their own *perspective* of statistical reasoning surrounding the task of comparing distributions. By listening to teacher *voice* we were able to gain further insight into our own understanding of variation as we worked to understand the teachers' reasoning. Although the literature clearly points to sampling distributions as a stumbling point for students in inferential statistics, we had thought that abundant experience with simulations involving sampling distributions within meaningful problems that would demonstrate their power would be sufficient to help teachers overcome this difficulty. In fact, the conflicts teachers had in using sampling distributions may have been compounded by the way in which sampling distributions and simulations were introduced together without providing sufficiently motivating tasks for teachers to create a need for them. We learned that a wealth of experience with sampling distributions to solve interesting problems was not sufficient for their understanding. We believe, given our analysis of teachers' reasoning in this area, that sampling distribution concepts need to be developed more slowly, allowing teachers to conceptually construct the notion of a sampling distribution rather than have it presented as part of a "good way" to solve the problem at hand.

Comparing distributions raises another important issue about variation—which variation are we referring to when we compare two distributions? With a single distribution, discussions of variation are meant to describe variation *within* the distribution at hand. Having two distributions to compare provides a motivation to compare variation *between* the distributions. For example, if we observe that the performance of males and females on a test differs by 2 points, what does this 2-point difference tell us? Could this difference just be due to random variation, or could it indicate a more meaningful underlying phenomenon? When comparing groups and considering variation between distributions, it is important to consider whether the data being compared is that of a *sample* or a *population.* Traditional introductory instruction in significance testing often uses sampling distributions as a way to generalize our findings from a sample to some larger, unknown population.

Whether data should be considered as a population or a sample is somewhat problematic in the context of a school's student assessment data and indicates that these distinctions are not always clear-cut (Chance, 2002). On one hand, it makes sense to consider a set of student test data from a school as its own population. When comparing two groups, however, sampling distributions can inform us as to whether the difference between groups is *meaningful,* hence pushing us to consider measures beyond descriptive statistics. Simulations can be used to support a broader, inference-like view of a difference even though we are not necessarily trying to generalize to a larger population. In this case, we can investigate the difference in means between male and female performance through the use of a randomization test. That is, under the null hypothesis that there is no difference between the performance of males and females on a test, if we were to randomize the students' genders and then compare the means of the two groups, how likely is a difference of 2 points to occur between males and females *just by chance*? On the other hand, we might want to conceptualize the two groups as samples in a larger population of all students who pass through a school over many years to make inferences about the school itself, even though the samples are not randomly selected, assuming one is willing to accept these as representative samples.

In working with teachers, we found that capturing and influencing teachers' statistical reasoning is much more complex than trying to understand and describe students' reasoning. Firstly, students are expected to be learners, but teachers consider themselves experts. Therefore, it is very difficult for most experienced teachers to admit what they do not know and be open to learning and discussing their reasoning. Fortunately, statistics is a content area in which few teachers are expected to have knowledge, making it a viable entrance for teachers to reexperience being learners. Secondly, unless experienced teachers are enrolled in a masters program, they are usually not an easily accessible group for the kind of long-term study that can affect teachers' thinking. The study described here began with an agreement between a school principal and our research group to commit the entire mathematics department of seven teachers to the research project, including a 2-week summer institute. By the end of the study however, only the two strongest of the seven original teachers remained. This raises both an important question and limitation of the study. First, how one can engage experienced secondary teachers in research that hopes to both influence and study teacher learning and practice? Second, the four teachers in the study likely had higher mathematical content knowledge than might be considered typical. In addition, they were very committed to improving their own practice, were highly engaged during activities and discussion, and were more open than most to consider weaknesses in their own understanding.

Comparing two groups provides a rich context in which to build statistical reasoning. At a very early age in school, group comparisons can provide an impetus to collect data and later, to view data as a distribution. At an advanced level, an interesting problem involving comparing distributions can stimulate learners to consider not only measures of dispersion within each group, but comparisons of measures between groups, and hence to consider variation within the measures themselves. Just as algebra and calculus are considered to be gatekeepers to higher mathematics, understanding sampling distributions may be a gatekeeper to advanced statistical reasoning. However, simply presenting sampling distributions as a precursor to hypothesis testing may aggravate the difficulty learners have with its underlying concepts.

Further work is needed in better understanding reasoning about sampling distributions as well as ways to think about facilitating learners' conceptual development of variation within a distribution with an eye toward developing a tolerance and expectation for variation in statistical measures. Understanding sampling distributions is by no means a cure for the difficulty of understanding variation of any sort, or toward loosening a deterministic view of statistics and data analysis. It is the authors' hope, however, that better understanding of teachers' reasoning about comparing groups will open further discussion of building an intuition of variation in data and statistics for teachers as well as students.

IMPLICATIONS

We ascertained that comparing distributions holds great potential for encouraging learners to broaden their view of statistics and data. As researchers, we found comparing distributions to be a fruitful arena for expanding teachers' understanding of distribution and conceptions of variability as well as a motivating reason to introduce sampling distributions. However, we found it important to specify which kind of variation we are discussing when comparing two distributions. Teachers' reasoning about variation in the context of group comparisons was examined in three areas: variation *within* a distribution, variation *between* groups (variation of measures), and the struggle to interpret the difference between these two types of variation. The importance of making this distinction surprised us, and motivated us to consider both our own understanding and the way in which we planned our conjectured learning trajectory. This study implies that sources of variation in both data and in measures need to be discussed frequently when working with data, and again as measures are compared between distributions, to engender a tolerance for variation both within and between distributions.

At a more advanced level of statistical content, our study supports the findings of delMas et al. (1999) about the difficulty in understanding sampling distributions and implies that the teaching of sampling distributions needs to be done more carefully. Furthermore, traditional teaching of hypothesis and significance testing and the overreliance on computer simulations may actually promote misconceptions rather than advance understanding of sampling distributions. In addition, discussion about the distinctions and ambiguities between considering data as a sample or a population need to occur in the teaching of significance testing and among the research community.

H. G. Wells predicted decades ago that "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write" (quoted in Snee, 1990, p. 117). If our goal is to promote statistical reasoning in our students, we must better understand and engender the statistical thinking and reasoning of teachers.

Snee (1990) highlights in his definition of statistical thinking in the quality control industry the importance of a recognition that "variation is all around us, present in everything we do" (p. 118). The concept of variation needs to be engendered early and continuously when teaching statistical reasoning. The teaching of statistics throughout schooling, with an emphasis on distribution and variation, may provide a way to loosen the deterministic stance of teachers, students, and the public toward data and statistics. More research is needed in this area.

REFERENCES

Abelson, R. (1995). *Statistics as principled argument.* Hillsdale, NJ: Erlbaum.

Chance, B. L. (2002). Personal communication (email: April 11, 2002).

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

Cohen, D. K., & Ball, D. L. (1990). Relations between policy and practice: A commentary. *Educational Evaluation and Policy Analysis, 12*(3), 249–256.

Confrey, J. (1991). Learning to listen: A student's understanding of powers of ten. In E. von Glasersfeld (Ed.), *Radical constructivism in mathematics education* (pp. 111–138). Dordrecht, The Netherlands: Kluwer Academic Publishers.

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: Cambridge University Press.

Confrey, J. (in preparation). *Systemic crossfire.* Unpublished manuscript.

Confrey, J., & Makar, K. (2002). *Developing secondary teachers' statistical inquiry through immersion in high-stakes accountability data.* Paper presented at the Twenty-fourth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education (PME-NA), Athens, GA.

delMas, R. C., Garfield, J., & Chance, B. L. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education, 7*(3).

Erickson, T. (2001). *Data in depth: Exploring mathematics with Fathom.* Emeryville, CA: Key Curriculum Press.

Finzer, W. (2000). Fathom (Version 1.1). Emeryville, CA: Key Curriculum Press.

Friel, S., Curcio, F., & Bright, G. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education, 32*(2), 124–158.

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

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.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998). *Connected Mathematics: Data about us.* White Plains, NY: Seymour.

Lehrer, R., & Schauble, L. (2000). Modeling in mathematics and science. In R. Glaser (Ed.), *Advances in instructional psychology: Educational design and cognitive science* (Vol. 5, pp. 101–159). Mahwah, NJ: Erlbaum.

Lieberman, A., & Wood, D. R. (2003). *Inside the National Writing Project: Connecting network learning and classroom teaching.* New York: Teachers College Press.

Makar, K., & Confrey, J. (2002). *Comparing two distributions: Investigating secondary teachers' statistical thinking.* Paper presented at the Sixth International Conference on Teaching Statistics (ICOTS-6), Cape Town, South Africa.

Meletiou, M. (2000). *Developing students' conceptions of variation: An untapped well in statistical reasoning.* Unpublished dissertation, University of Texas, Austin.

National Writing Project. (2002, April). *National Writing Project Mission.* Author. Retrieved April 28, 2002, from www.writingproject.org

QSR. (1999). NVivo (Version 1.1). Melbourne, Australia: Qualitative Solutions and Research Pty. Ltd.

Snee, R. (1990). Statistical thinking and its contribution to total quality. *The American Statistician, 44*(2), 116–121.

Stake, R. E. (1994). Case studies. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research.* Thousand Oaks, CA: Sage.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory.* Thousand Oaks, CA: Sage.

TERC. (1998). *Investigations in number, data, and space.* White Plains, NY: Seymour.

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

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