

Enhancing Prospective Teachers' Coordination of Center and Spread: A Window into Teacher Education Material Development¹

Hollylynn S. Lee & J. Todd Lee

This paper describes a development and evaluation process used to create teacher education materials that help prepare middle and secondary mathematics teachers to teach data analysis and probability concepts with technology tools. One aspect of statistical reasoning needed for teaching is the ability to coordinate understandings of center and spread. The materials attempt to foster such coordination by emphasizing reasoning about intervals of data rather than a single focus on a point estimate (e.g., measure of center). We take a close look at several different data sources across multiple implementation semesters to examine prospective mathematics teachers' ability to reason with center and spread in a coordinated way. We also look at the prospective teachers' ability to apply their understandings in pedagogical tasks. Our analysis illustrates the difficulty in both achieving this understanding and transferring it to teaching practices. We provide examples of how results were used to revise the materials and address issues of implementation by mathematics teacher educators.

Data analysis, statistics, and probability are becoming more important components in middle and high school mathematics curricula (National Council of Teachers of Mathematics, 2000; Franklin et al., 2005). Therefore, university teacher educators are challenged with how to best prepare prospective mathematics teachers to teach these concepts. The challenge is exacerbated by the fact that many of these prospective teachers have not had meaningful opportunities to develop an understanding of pivotal statistical and probabilistic ideas (e.g., Stohl, 2005). Although simulation and data analysis tools—graphing calculators, spreadsheets, *Fathom*, *TinkerPlots*, *Probability Explorer*—may be available in K-12 classrooms, there is a need for high quality teacher education curriculum materials. Such curriculum materials can help teacher educators become comfortable with and incorporate tools for teaching

probability and data analysis. These teacher education curricula need to primarily aim for prospective teachers to develop a specific type of knowledge related to statistics that includes a deeper understanding of: (a) data analysis and probability concepts, (b) technology tools that can be used to study those concepts, and (c) pedagogical issues that arise when teaching students these concepts using technology (Lee & Hollebrands, 2008b; Lesser & Groth, 2008).

The authors of this paper are part of a team engaged in a teacher education materials development project, funded by the National Science Foundation, to create units of course materials—modules with about 18-20 hours of class materials with additional assignments—to integrate technology and pedagogy instruction in various mathematical contexts. The project intends to create three modules that could be distributed separately and used in mathematics education methods courses, mathematics or statistics content courses for teachers, or professional development workshops focused on using technology to teach mathematics and statistics. The modules are not designed for teachers to use directly with their students. Rather, the developers anticipate that after using the materials teachers will have the knowledge needed to create their own technology-based activities. The three modules will focus on the teaching and

Dr. Hollylynn Stohl Lee is an Associate Professor of Mathematics Education at North Carolina State University. Her research interests include the teaching and learning of probability and statistics with technology.

Dr. J. Todd Lee is a Professor of Mathematics at Elon University. He is interested in undergraduate mathematics education, including the probability and statistics learning of pre-service teachers.

¹ The work on this curriculum development and research was supported by the National Science Foundation under Grant No. DUE 0442319 and DUE 0817253 awarded to North Carolina State University. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily reflect the views of the National Science Foundation. More information about the project and materials can be found at <http://ptmt.fi.ncsu.edu>.

learning of data analysis and probability, geometry, and algebra.

The first module focuses on learning to teach data analysis and probability with technology tools, including TinkerPlots, Fathom, spreadsheets, and graphing calculators (Lee, Hollebrands, & Wilson, 2010). This module is designed to support a broad audience of prospective secondary teachers. For many prospective teachers, engaging in statistical thinking is a different process than that which they have been engaged in teaching and learning mathematics (e.g., delMas, 2004). Thus, it is important to engage prospective teachers as active learners and doers of statistics. The module incorporates several big ideas that can support teachers as they learn to teach data analysis and probability: engaging in exploratory data analysis; attending to distributions; conceptually coordinating center and spread in data and probability contexts; and developing an understanding of, and disposition towards, statistical thinking as different from mathematical thinking. For this paper, we focus solely on one of these big ideas as we discuss the material development process using the following guiding question: How can we use technology tools to enhance prospective mathematics teachers' coordination of center and spread? We analyzed several forms of data to revise the teacher education materials. The results provide insight into ways prospective mathematics teachers may reason about center and spread in a coordinated way.

Why Focus on Coordinating Center and Spread?

Coordinating measures of center and spread has been identified as a central reasoning process for engaging in statistical reasoning (e.g., Friel, O'Connor, & Mamer, 2006; Garfield, 2002; Shaughnessy, 2006). In particular, Garfield (2002) noted that part of reasoning about statistical measures is "knowing why a good summary of data includes a measure of center as well as a measure of spread and why summaries of center and spread can be useful for comparing data sets" (Types of Correct and Incorrect Statistical Reasoning section, para. 11).

Single-point indicators, used as a center of a distribution of data (e.g., mean or median) or as an expected value of a probability distribution, have been over-privileged in both mathematics curricula (Shaughnessy, 2006) and statistical research methods (Capraro, 2004). When used with samples, single-point central indicators may not be accurate signals of what is likely an underlying noisy process (Konold & Pollatsek, 2002). Many others argue that attending to variation is critical to developing an understanding of

samples and sampling distributions (e.g., Franklin et al, 2005; Reading & Shaughnessy, 2004; Saldanha & Thompson, 2002; Shaughnessy, 2006).

Understanding variability, both within a single sample and across multiple samples, can be fostered through attending to intervals: Intervals embody both central tendency and spread of a data set (Reading & Shaughnessy, 2004). Attending to intervals aligns well with the many voices of concern in professional communities on the limitation of null hypothesis significance testing, which rely on single-point p -values. For example, the medical industry has taken major moves toward examining and reporting data through alternative tools, confidence intervals being foremost (Gardner & Altman, 1986; International Committee of Medical Journal Editors, 1997). Other areas, such as psychology, ecology, and research in mathematics education, are also moving in this direction (Capraro, 2004; Fidler, 2006).

When describing expected outcomes of a random process, interval thinking can make for a powerful, informative paradigm shift away from single-point estimates. Statistics education researchers have advocated this shift in focus (e.g., Reading & Shaughnessy, 2000, 2004; Watson, Callingham, & Kelly, 2007). For example, in a fair coin context, describing the number of heads that may occur when tossing a coin 30 times is better described as "typically about 12 to 18 heads" rather than "we expect 15 heads." The latter statement does not acknowledge the variation that could occur. As Reading and Shaughnessy (2000, 2004) have noted, many students will initially provide single point values in tasks asking for expectations from a random process, but this is likely related to the common use of such questions as "'What is the is the probability that ...?' Probability questions just beg students to provide a point-value response and thus tend to mask the issue of the variation that can occur if experiments are repeated" (p. 208, Reading & Shaughnessy, 2004). Explicitly asking for an interval estimate may illicit a classroom conversation that focuses students' attention on variation.

Prospective and practicing teachers have demonstrated difficulties similar to middle and high school aged students in the following areas: considering spread of a data set as related to a measure of center (Makar & Confrey, 2005), appropriately accounting for variation from an expected value (Leavy, 2010), and a tendency to have single-point value expectations in probability contexts (Canada, 2006). Thus, there is evidence to suggest mathematics

educators should help prospective teachers develop an understanding of center and spread that can allow them and their students to reason appropriately about intervals in data and chance contexts. The aim of our materials development and evaluation efforts reported in this paper is to document one attempt to foster such reasoning and to reflect upon how the evaluative results informed improving the materials and suggestions for future research.

Design Elements in Data Analysis and Probability Module

From 2005 to 2009, the Data Analysis and Probability module materials for prospective secondary and middle mathematics teachers were developed, piloted, and revised several times. To facilitate understanding of measures of center and spread in a coordinated way, Lee et al. (2010) attempted to do the following:

1. Emphasize the theme of center and spread throughout each chapter in the material, with the coordination between the two explicitly discussed and emphasized through focused questions covering both content and pedagogical issues.
2. Use dynamic technology tools to explore this coordination.
3. Place the preference for intervals above that of single-point values even if the construction of these intervals is reliant upon measures of center and spread.

Lee et al., with consultation from the advisory board and a content expert, attempted to attend to these elements, along with other design elements aimed at developing prospective teachers' understanding of data analysis and probability, technology issues, and appropriate pedagogical strategies. A discussion of the design of the entire module as it focuses on developing technological pedagogical content knowledge for statistics is discussed in Lee and Hollebrands (2008a, 2008b).

Methods

The project team followed curricular design and research method cycles as proposed by Clements (2007), including many iterations of classroom field-testing with prospective teachers, analysis of field-testing data, and subsequent revisions to materials. Our primary research site, a university in the Southeast region of the US, has consistently implemented the module in a course focused on teaching mathematics with technology serving third- and fourth-year middle

and secondary prospective teachers and beginning graduate students who need experience using technology. A typical class has between 13 and 19 students. In Fall 2005, during the five-week data analysis and probability module, the instructor used the pre-existing curriculum for the course to serve as a comparison group to the subsequent semesters. The students took a pretest and posttest designed to assess content, pedagogical, and technology knowledge related to data analysis and probability.

In each of the subsequent semesters from 2006-2007, the same instructor as in Fall 2005 taught a draft of the five-week Data Analysis and Probability module from our textbook (Lee et al., 2010) with a request that the curriculum be followed as closely as possible. In addition, the module was implemented in a section of the course taught by a different instructor, one of the authors of the textbook, in Spring 2007. During the first two semesters of implementation, class sessions were videotaped and several students were interviewed. In the first three semesters of implementation, written work was collected from students and pre- and post-tests were given. Since 2007, many other instructors have used the materials at institutions across the US and improvements and slight modifications were made based on instructor and student feedback, with final publication in 2010 (Lee et al.).

For this study, we are using several sources of data for our analysis of how prospective teachers may be developing a conceptual coordination between center and spread in data and probability contexts, with a particular focus on interval reasoning. Our data sources include: (a) examples of text material from the module, (b) a video episode from the first semester of implementation in which prospective teachers are discussing tasks concerning probability simulations, (c) prospective teachers' work on a pedagogical task, and (d) results from the content questions on the pre- and post-tests across the comparison and implementation semesters through Spring 2007.

Analysis and Results

We discuss the analysis and results according to the four data sources we examined. In each section we describe the analysis processes used and the associated results.

Emphasis in Materials: Opportunities to Learn

To begin our analysis, we closely examined the most recent version of the text materials for opportunities for prospective teachers to develop a coordinated conceptualization between center and

spread. The materials begin by helping prospective teachers informally build and understand measures of center and spread in the context of comparing distributions of data (Chapter 1) and then explore a video of how middle grades students compare distributions (Chapter 2).

In Chapter 3, prospective teachers consider more deeply how deviations from a mean are used to compute measures such as variation and standard deviation. In Chapter 4, the materials build from this notion in a univariate context to help students consider measures of variation in a bivariate context when modeling with a least squares line. The focus on spread and useful intervals in a distribution continues in Chapters 5 and 6 where prospective teachers are asked to describe distributions of data collected from simulations, particularly attending to variation from expected values within a sample, and variation of results across samples. These last two chapters help prospective teachers realize that smaller sample sizes are more likely to have results that vary considerably from expected outcomes, while larger sample sizes tend to decrease this observed variation.

We only considered places in the text materials where the authors had made an explicit reference to these concepts in a coordinated way as opportunities for prospective teachers to develop a conceptualization of coordinating center and spread. We closely examined the text materials to identify instances where there was an *explicit* emphasis placed on coordinating center and spread in (a) the written text and technology screenshots, (b) content and technology tasks, and (c) pedagogical tasks. One researcher initially coded each instance throughout the textbook, the researchers then

conferred about each coded instance to ensure that both agreed that an instance was legitimate. We tallied the final agreed-upon instances in each chapter as displayed in Table 1. We also specifically marked those instances addressing coordinating center and spread that placed special emphasis on promoting interval reasoning as displayed in Table 1. For an example of instances coded as focused on interval reasoning, see Table 2. The point of this content analysis was to identify where and how often the authors of the materials had actually provided opportunities for prospective teachers to coordinate center and spread and engage in reasoning about intervals. This analysis could also point out apparent gaps where opportunities may have been missed to the author team.

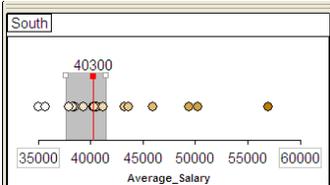
As seen in Table 1, every chapter contained content and technology tasks as well pedagogical tasks that emphasized the coordination of center and spread. This coordination was discussed in the text along with any diagrams and technology screenshots in all but Chapter 2 (which is a video case with minimal text), with slightly heavier emphases in Chapters 4 and 5. Chapters 5 and 6 have the most content and technology tasks focused on coordinating center and spread. Of particular importance is that an explicit focus on interval reasoning only appears in Chapter 1, 5, and 6, with Chapter 5 containing a particularly strong emphasis. Although evidence suggests the design of the materials provides opportunities to build understanding of center and spread throughout, attention to this in the early versions of the materials is uneven, particularly in terms of emphasizing interval reasoning.

Table 1
Instances in Module of Coordinating Center and Spread

	Instances of coordinating center and spread			Percent of instances with focus on interval reasoning
	Text	Content & technology. task	Pedagogical task	
Ch 1: Center, Spread, & Comparing Data Sets	3	5	2	50%
Ch 2: Analyzing Students' Comparison of Two Distributions using <i>TinkerPlots</i>	0	2	2	0%
Ch 3: Analyzing Data with <i>Fathom</i>	2	5	3	0%
Ch 4: Analyzing Bivariate Data with <i>Fathom</i>	5	3	3	0%
Ch 5: Designing and Using Probability Simulations	4	13	4	76%
Ch 6: Using Data Analysis and Probability Simulations to Investigate Male Birth Ratios	1	15	1	59%

Table 2

Examples of Instances in Materials Coded as Opportunities to Coordinate Center and Spread and Promote Interval Reasoning

Written text and screenshots	Content and technology tasks	Pedagogical tasks
<p>Students may attend to clumps and gaps in the distribution or may notice elements of symmetry and peaks. Students often intuitively think of a “typical” or “average” observation as one that falls within a modal clump... Use the divider tool to mark off an interval on the graph where the data appear to be clumped.</p> <p>(Chapter 1, Section 3, p. 11)</p>	<p>Q17: Use the Divider tool and the Reference tool to highlight a clump of data that is “typical” and a particular value that seems to represent a “typical” salary. Justify why your clump and value are typical. (Chapter 1, Section 3, p. 13)</p> 	<p>Q19: How can the use of the dividers to partition the data set into separate regions be useful for students in analyzing the spread, center and shape of a distribution? (Chapter 1, Section 3, p. 14)</p>
<p>In our context, we are interested in how much the proportion of freshmen returning to Chowan College will vary from the expected 50%. To examine variation from an expected proportion, it is useful to consider an interval around 50% that contains most of the sample proportions.</p> <p>(Chapter 5, Section 3, p.102)</p>	<p>Q11. Given a 50% estimate for the probability of retention, out of 500 freshmen, what is a reasonable interval for the proportion of freshmen you would expect to return the following year? Defend your expectation. (Chapter 5, Section 3, p. 100)</p> <p>Q16. If we <i>reduced</i> the number of trials to 200 freshmen, what do you anticipate would happen to the interval of proportions from the empirical data around the theoretical probability of 50%? Why? Conduct a few samples with 200 trials and compare your results with what you anticipated. (Chapter 5, Section 3, p. 103)</p>	<p>Q19. Discuss why it might be beneficial to have students simulate the freshman retention problem for several samples of sample size 500, as well as sample sizes of 200 and 999. (Chapter 5, Section 3, p. 103) [Implied emphasis on interval reasoning because it is one of the follow-up questions to Q16.]</p>

Classroom Episode from Chapter 5

Because Chapter 5 contained the largest focus on coordinating center and spread via interval reasoning, we analyzed a 2.5 hour session of a class engaging in Chapter 5 material from the first implementation cycle. The researchers viewed the class video several times and critical episodes (Powell, Francisco, & Maher, 2003) were identified as those where prospective teachers or the teacher educator were discussing something that had been coded as an “instance” in Chapter 5 as seen in Table 1. Each critical episode was then more closely viewed to examine how the reasoning being verbalized by prospective teachers or the teacher educator indicated an understanding of coordinating center and spread and the use of interval reasoning.

It is not possible to present a detailed analysis of the entire session; however we present classroom discussions around several of the interval reasoning

tasks shown in Table 2. Consider the following question posed in the text materials:

Q11: Given a 50% estimate for the probability of retention, out of 500 freshmen, what is a reasonable interval for the proportion of freshmen you would expect to return the following year? Defend your expectation.

This question follows material on the technical aspects of using technology to run simple simulations and how to use these simulations as a model for real world situations. Immediately prior to Question 11 prospective teachers are asked to write (but not run) the commands needed on a graphing calculator that would run multiple simulations of this scenario. In answering Question 11, several prospective teachers propose three intervals they considered to be reasonable for how many freshmen out of 500 they expect to return the following year at a college with a 50% retention rate; 230-270, 225-275, and 175-325. The teacher educator asked a prospective teacher to explain his reasoning for

the interval 230-270. (T denotes teacher educator and PT denotes a prospective teacher)

- T: Can you tell me why you widened the range?
- PT1: I didn't, I narrowed it
- T: Tell me why you narrowed it
- PT1: 500 is a big number. So I thought it might be close to 50%.
- T: So you thought because 500 is a big number it would be closer to
- PT1: Half
- T: To half, closer to 50%. So, MPT1 [who proposed an interval of 175-325], why did you widen the range? This [pointing to 225-275 on board] was the first one thrown out, why did you make it bigger?
- PT2: Well it's all according to how long you're going to do the simulation.
- T: Out of 500 students how many [slight pause] what range of students will return? Do you think it will be exactly 50% return?
- PT2: Probably not
- T: So for any given year, what range of students might return, if you have 500 for ever year?
- PT2: 175 to 325
- T: Ok. So can you tell me why?
- PT2: Without knowing anything I wouldn't go to a tight range.
- T: Because you don't have enough information.
- PT3: It's like the coin flips; you have some high and some low, so it might not fall into the 225 to 275 interval.
- PT4: I'd say it will most likely fall into that first range, but it's not a bad idea to be safe and say it can go either way.

First, all intervals were given in frequencies, rather than proportions. This is likely an artifact of the wording of Question 11 during that implementation cycle. In that version of the materials, the question did not specifically use the word proportion. All intervals

suggested by the prospective teachers are symmetric around an expected retention of 250 (50%) of 500 freshmen. Two of the intervals have widths less than 10% of the range, or a maximum variation of 5% from the mean, while the largest proposed interval 175-325 suggests a variation of $\pm 15\%$. The smaller intervals have around 93% and 98% chances of containing the future retention proportion, while the largest interval will succeed with an almost mathematical certainty. While one prospective teacher reasoned that 500 is a large enough sample to expect values "close" to 50%, another is much more tentative and casted a wider net due to an uncertainty about the number of times the simulation would be run. This prospective teacher, and the two that responded afterward, may be trying to capture all possible values, rather than consider a reasonable interval that would capture most values. Or they may merely be dealing with the difficulties of estimating the binomial distribution of 500 trials. Only one prospective teacher justified an interval by explicitly reasoning from an expected value, and there were no justifications. The teacher educator did not question why the intervals were symmetric about the expected value. The reasoning of the prospective teacher is similar to that noticed by Canada (2006) in his research with prospective elementary teachers. Canada noted, "almost all of my subjects pointed out that more samples would widen the overall range, while very few subjects suggested that more samples would also tighten the subrange capturing most of the results" (p. 44).

After about 30 minutes of exploration using a calculator to run simulations, the teacher educator asked each prospective teacher to run two simulations of the "50% retention rate of 500 freshmen" and compute the proportion of freshmen returning. The teacher educator collected and displayed this data as a dot plot in *Fathom* (Figure 1). This is the second time during this lesson the teacher educator used *Fathom* to collect data from individual's samples and display them as a distribution. This teacher educator's move was not suggested in the curriculum materials; however its value in indicating a public record and display of pooled class data is duly noted and used in revisions to suggest such a way to display class data in aggregate form.

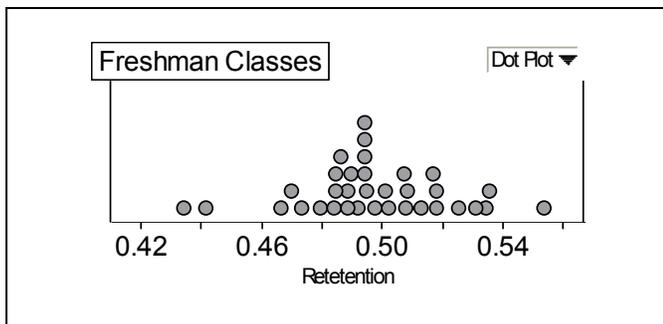


Figure 1. Distribution of 34 sample proportions pooled from class and displayed.

The plot in Figure 1 appears quite typical for what might occur with 34 samples of 500, with a modal clump between 0.48 and 0.51. The teacher recalls the predicted intervals and asks:

T: If we take a look at the distribution of this data in a graph [displays distribution in Figure 1], is that kind of what you would assume? We ran the simulation of 500 freshman 34 times. So we notice, we assumed 50%. Are we around 50%? How many times are we at 50%?

PT: One

T: Here are your predictions from earlier on the number of students you might see in a range [three proposed intervals]. Our proportion range is about from 0.44 to 0.53. Think any of these ranges for the students are too wide or too narrow...?

The teacher educator immediately drew attention to the expected value of 50% and variation from that expectation with comments of “around 50%” and “at 50%.” The conversation shifted as the teacher educator appeared to draw their attention to the entire range of proportion values, rather than on a modal clump around the expected value. It appears that both the teacher educator and the prospective teachers interpreted the request for a “reasonable interval” in the textbook question to mean the range of all sample proportions likely to occur, or that do occur.

The discussion continued as the teacher educator had the prospective teachers use an algorithm to convert the proportion range, which was re-estimated as 0.43-0.55, to frequencies 215-275 so they could compare the predicted intervals. They noted the similarity of the sample range to two of the proposed

intervals, and they noted that the range is not symmetric around 0.5 and therefore is “not like we thought” [FPT1]. The teacher educator then focused the class back on the expected value of 50% and asked why they did not get more samples with a retention of 50%. One prospective teacher offered a reason related to a low sample size and another suggested the graphing calculator’s programming may be flawed. Another prospective teacher countered the idea:

PT: If it [graphing calculator] is programmed to act randomly, it is not going to recognize any particular value. And it will..., point 5 is the theoretical value. But the actual values don’t have to be point 5, they should be close to point 5, which most of them are.

The teacher educator did not pursue the conversation about the graphing calculator, but instead asked a question based on Question 16, as seen in Table 2, and two questions that follow in the text. We will use this conversation to consider how students reason about the relationship between sample size and variation from the expected center.

T: So let’s say instead of doing 500 freshmen, we would decrease this set to 200. How do you think the range might differ, or if we increased to 999 how might the range of proportions be different?

PT: It would be narrower.

T: Narrower for which way, if we reduced to 200 or increased to 999?

PT: 999

T: Why do you think it would be narrower?

PT: The more trials there are, the closer it will be to the true mean.

T: [Asks students if they agree, about half the class raise their hand.]

... [Other prospective teachers make similar comments.]

T: If we decrease to 200 trials in each sample from 500 do you expect the range to be similar or do you expect it to be wider or narrower or similar??

PT: Wider. With a smaller sample you will have more variability.

T: So you are going with the idea that a smaller sample will have more variability. Does everyone agree or disagree? [many prospective teachers say agree].

This episode suggests that at least some prospective teachers were developing an understanding of the relationship between the freshman class size and the variation in the distribution of sample proportions from repeated samples. This suggests that although they may have not initially approached the task with an expectation of an appropriate interval for what might be typical, many came to reason, through the extended activity and repeated simulations, that the reasonable interval widths were affected by sample size. This again aligns with Canada's (2006) result that his instructional intervention helped more of the prospective elementary teachers consider the role of sample size as an influence on the variation of results around the expected value.

It seems as though explicitly asking about intervals provided opportunities for class discussions that went beyond the discussion of a single expected value, in this case 50%. Such an opportunity can help develop the notion that with random processes comes variation, and that understanding how things vary can be developed through reasoning about intervals rather than merely point-estimates of an expected center value. However, symmetry may well have been strongly used due to the retention rate being 50%; it may be beneficial to incorporate an additional question using retention rates other than 50%.

Pedagogical Task Following Chapter 5

The ultimate goal of these materials is to develop prospective teachers' abilities to design and implement data analysis and probability lessons that take advantage of technology. Fortunately, there are many opportunities within the materials to engage in pedagogical tasks. One such task followed the previously described prospective teachers' work in Chapter 5. As a follow-up to our examination of the classroom interactions for Chapter 5, we examined how these same prospective teachers may have applied their developing understandings in a pedagogical situation. The task describes a context in which college students are able to randomly select from three gifts at a college bookstore and then asks:

Explain how you would help students use either the graphing calculator, Excel, or Probability Explorer to simulate this context. Explicitly describe what the commands

represent and how the students should interpret the results. Justify your choice of technology.

Of particular interest to us was whether prospective teachers would plan to engage their students in using large sample sizes, using repeated sampling, and using proportions rather than frequencies to report data. We also were interested in whether they would promote or favor interval reasoning in lieu of point-value estimates.

Each prospective teacher submitted a written response to this task. Seventeen documents were available for analysis. Each response was summarized with respect to several categories: (a) which technology was chosen and why, (b) how the tool would generally be used, (c) what use was made of sampling and sample size, (d) how representations for empirical data would be used, and (e) what they want students to focus on in their interpretation. The summaries were used to identify patterns across cases as well as interesting cases.

The majority chose to use a graphing calculator (10 of 17), only 5 of the 17 prospective teachers planned experiences for their students that incorporated repeated samples, and only 7 used proportions. In addition, 10 prospective teachers focused explicitly on a point estimate, one used both a point and interval estimate for interpreting a probability, while six of the responses to the task were not explicit enough to tell what the prospective teacher intended. Thus, the majority planned for students to simulate one sample (sample sizes vary across lessons, but many were less than 50) and to make a point estimate of the probability from that sample.

The prospective teachers did not provide much evidence, during the week immediately following their discussion of the material in Chapter 5, that they were able to transfer their developing understandings of interval reasoning in a probability context to a pedagogical situation. It seems that, for most, any progress made during the class discussions did not have a transference effect into their pedagogy.

Pre- and Post- Tests

Pre- and post-tests were used to create a quantitative measure that might indicate prospective teachers' conceptual changes. The 20 questions comprising the content section of the pre- and post-assessment were selected from Garfield (2003) and other items from the ARTIST database (<http://app.gen.umn.edu/artist/index>). These items assess general statistical reasoning concerning concepts

included in the text materials (e.g., coordinating center and spread, interpreting box plots, interpreting regression results and correlations). These questions were administered to the prospective teachers both before and after the Data Analysis and Probability

module, and the scores were combined pair-wise as normalized gains. By normalized gains, we mean the percentage increase of a student's available advancement from the pre- to post-test (Hake, 1998).

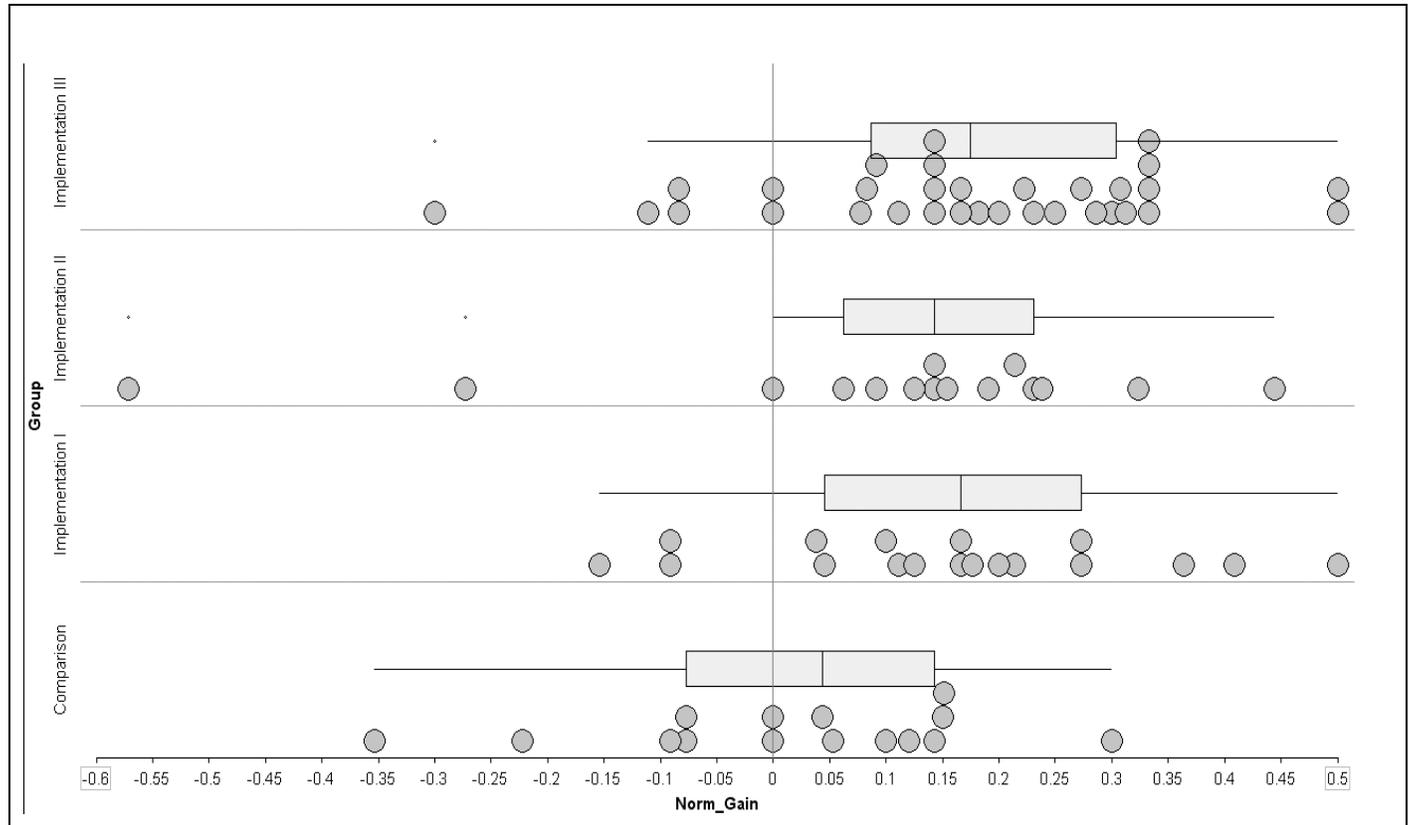


Figure 2. Distribution of normalized gain scores for each group of prospective teachers.

The Comparison group ($n=15$) plot shows normalized gains realized in Fall 2005 using the traditional curricula for the course, prior to implementation of the new materials. Compared against this group are the normalized gains from three different semesters (four total sections) in which the materials were implemented. There were major revisions to the text materials between Implementation I ($n = 18$) and II ($n = 15$), but only minor edits before Implementation III ($n = 32$, based on two sections). However, prospective teachers in the Implementation III group were the first that used the module as a textbook for reference in and out of class. Other than exposure to different curricula, it seems reasonable to assume that the prospective teachers across all sections came from the same population.

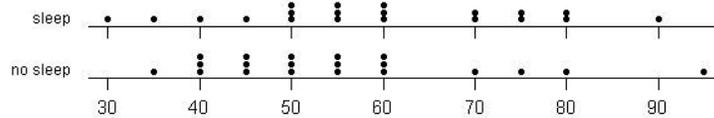
Visual inspection reveals a distinct increase in gains in the implementation groups with respect to the

comparison group. The gains seem to translate by more than 0.10, but we see little change in the amount of variation in the inter-quartile ranges. This assessment is in agreement with Monte Carlo permutation tests, $n = 50,000$, comparing both means, $p = .009$, and medians, $p = .006$, of the comparison group with those of the pooled implementations. However, comparing gains across the whole test is not part of our current focus in this paper.

Looking at the normalized gain scores for the entire content subsection of the test obscures the performance on particular questions. Thus, we selected and closely examined four questions from the test that address various aspects of our focus on the coordination of center and spread and the alternative use of intervals (see Figure 3). In Table 3, we record the percentage of students who answered the multiple choice questions correctly on the pre- and post-test.

3. The Springfield Meteorological Center wanted to determine the accuracy of their weather forecasts. They searched the records for those days when the forecaster had reported a 70% chance of rain. They compared these forecasts to records of whether or not it actually rained on those particular days. The forecast of 70% chance of rain can be considered *very* accurate if it rained on:
- 95% - 100% of those days.
 - 85% - 94% of those days.
 - 75% - 84% of those days.
 - 65% - 74% of those days.**
 - 55% - 64% of those days.
10. Half of all newborns are girls and half are boys. Hospital A records an average of 50 births a day. Hospital B records an average of 10 births a day. On a particular day, which hospital is more likely to record 80% or more female births?
- Hospital A (with 50 births a day)
 - Hospital B (with 10 births a day)**
 - The two hospitals are equally likely to record such an event.

11. Forty college students participated in a study of the effect of sleep on test scores. Twenty of the students volunteered to stay up all night studying the night before the test (no-sleep group). The other 20 students (the control group) went to bed by 11:00 pm on the evening before the test. The test scores for each group are shown on the graph below. Each dot on the graph represents a particular student's score. For example, the two dots above 80 in the bottom graph indicate that two students in the sleep group scored 80 on the test.



Examine the two graphs carefully. From the 6 possible conclusions listed below, *choose the one with which you most agree*.

- The no-sleep group did better because none of these students scored below 35 and a student in this group achieved the highest score.
- The no-sleep group did better because its average appears to be a little higher than the average of the sleep group.
- There is no difference between the two groups because the range in both groups is the same.
- There is little difference between the two groups because the difference between their averages is small compared to the amount of variation in the scores.**
- The sleep group did better because more students in this group scored 80 or above.
- The sleep group did better because its average appears to be a little higher than the average of the no-sleep group.

15. Each student in a class tossed a penny 50 times and counted the number of heads. Suppose four different classes produce graphs for the results of their experiment. There is a rumor that in some classes, the students just made up the results of tossing a coin 50 times without actually doing the experiment. *Please select each* of the following graphs you believe represents data from actual experiments of flipping a coin 50 times.

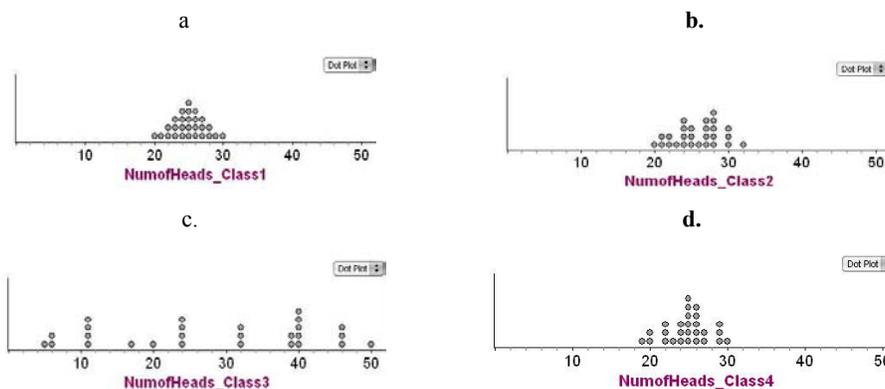


Figure 3. Sample pre- and post-test questions on center, spread, intervals, and variability.

Table 3
Correct Response Rates on Four Test Questions.

Question	Correct Answer	Comparison <i>n</i> = 15		Implementation I <i>n</i> = 18		Implementation II <i>n</i> = 15		Implementation III <i>n</i> = 32	
		Pre	Post	Pre	Post	Pre	Post	Pre	Post
3	d	47%	47%	44%	50%	53%	53%	53%	53%
10	b	40%	80%	44%	89%	33%	80%	38%	66%
11	d	53%	20%	11%	22%	33%	20%	25%	25%
15	b & d	47%	40%	56%	67%	40%	33%	41%	56%

Across all implementation semesters and the comparison group, prospective teachers made little to no improvement in their ability to interpret the accuracy of a 70% probability in data as an interval around 70% (Question 3, answer d), with only about half of them correctly choosing the interval. Across all semesters, there was also little change in prospective teachers' ability to recognize the two reasonable distributions for a distribution of outcomes from repeated samples of 50 coin tosses (Question 15, answers b and d). As shown in response to Question 10 (answer b), prospective teachers appeared to improve their ability to recognize sampling variability with respect to sample size: They typically became more likely to recognize that Hospital B, with the smaller sample size, had a higher probability of having a percent of female births much higher (80%) than an expected 50%. Because the comparison group made similar gains on Question 10 as those who had engaged in using the new materials, it appears that merely engaging in learning about data analysis and probability may be helpful in one's ability to correctly respond to that question, regardless of curriculum material.

For Question 11, there was very little change in the percent of prospective teachers who correctly chose *d* to indicate that there was little difference between the groups with respect to center and the large spread, and in fact most chose *f*, a comparison done only on a measure of center. It is disappointing that more prospective teachers did not demonstrate a coordination of center and spread with this task on the posttest. It is interesting that in the Comparison group, about half initially reasoned correctly but that after instruction the majority chose to make a comparison based only on a measure of center (see Figure 3).

Perhaps the traditional curriculum placed a greater emphasis on measures of center and decision-making based on point estimates.

The main lesson we take from examining these pre- and post-test questions is that our materials, as implemented in 2006-2007, did not appear to substantially help prospective teachers improve their reasoning about center, spread, and intervals. For although we realized gains in the overall scores on statistical reasoning, a close look at four questions demonstrates little change.

Discussion

How do these results help answer our question about the task of developing prospective teachers' ability to use a coordinated view of center and spread? One design element used by Lee et al. (2010) was the deliberate and consistent focus on the coordination of center and spread. The module covers a broad range of material, written by three authors through many iterations and reviews from external advisors. Though the theme of coordination was maintained throughout the material, the emphasis was found to be quite inconsistent across chapters in an early version of the materials. Even more sporadic was the preference of intervals over point values with half the chapters excluding this theme. Even though the focus on intervals and modal clumping was consistent in the probability/simulation chapters, a few of the relevant test questions did not indicate any gains beyond those from general exposure to data and probability. To ascertain if these themes can strengthen the intuitions of clumping over point-value intuitions, the message must be reemphasized throughout the material.

Prospective Teachers' Developing Understandings

Developing a coordinated view of center and spread, or expectation and variation, as others have called it (e.g., Watson et al., 2007), is difficult. Watson and her colleagues found that hardly any students from ages 8 to 14 used reasoning that illustrated a coordinated perspective on expectation and variation in interview settings. Although Canada's (2006) prospective teachers made gains during his course in reasoning about intervals, it was not uncommon for the teachers to still give single point estimates as expected values. If students have difficulty in coordinating center and spread, then it is important for both prospective and in-service teachers to work towards developing their own coordinated views in data and chance settings.

There are not many studies that follow the development of prospective teachers' understandings of statistical ideas into teaching practices. Batanero, Godina, and Roa (2004) found that even when gains in content knowledge were made during instruction on probability, prospective teachers still prepared lesson plans that varied greatly in their attention to important concepts in probability. Lee and Mojica (2008) reported that practicing middle school teachers, in a course on teaching probability and statistics, exhibited inconsistent understandings of probability ideas from lessons in their classrooms. Thus, it is not surprising that in such a short time period the prospective teachers in our study did not develop their own understandings in ways they could enact in pedagogical situations. Leavy (2010) noted that a major challenge in statistics education of prospective teachers is "the transformation of subject matter content knowledge into pedagogical content knowledge" (p. 49). Leavy also noted in her study that prospective teachers who were able to demonstrate a reasonably strong understanding of informal inference, including accounting for variation from expected outcomes, had difficulties applying this knowledge to create informal inference tasks to use with their own students.

Informing Revisions to Materials

In accordance with curriculum development and research recommendations by Clements (2007), the results discussed in this paper informed the next iteration of revisions to the materials. Several questions were revised throughout the text and additional discussion points were inserted to help emphasize the coordination of center and spread and to provide additional opportunities for interval reasoning. For example, a major change occurred in Chapter 1 with

regard to the focus on interval reasoning. Consider the original questions on the left side of Table 4 with those on the right. Fall 2007 Q17 asks prospective teachers to simultaneously consider spread and center through use of the divider and reference tools in TinkerPlots. However, in recent revisions, the series of questions was recast and developed into a series that first has the prospective teachers consider intervals of interest in the upper 50%, middle 50%, and then something they deem to be a cluster containing many data points, i.e., a modal clump. After the experience with intervals, they are asked to use the reference tool to mark a point estimate they would consider a "typical" value and to reason how the shaded interval might have assisted them. This series of questions puts much more explicit attention on valuing intervals when describing a distribution. The authors also added Q25, which explicitly asks prospective teachers to consider how the use a specific technology feature (dividers) can assist students' reasoning.

Other revisions made throughout the chapters included minor wording changes that could shift the focus of attention in answering the question. For example the Fall 2007 version of Chapter 3 posed the question:

Q9. By only examining the graphs, what would you characterize as a typical City mpg for these automobiles?

This question was revised:

*Q9. By only examining the graphs, what would you characterize as a typical **range** of City mpg for these automobiles? [bolding added]*

Informing Support for Faculty

Making changes in the text material is not sufficient. Fidelity of implementation is important for ensuring prospective teachers have opportunities to attend to and discuss the major ideas in the materials. The big statistical ideas in the text (e.g., exploratory data analysis, distributions, variation, and coordinating center and spread) need to be made explicit to the course instructor through different avenues, such as a facilitator's guide or faculty professional development. Such a guide has been developed and is available at <http://ptmt.fi.ncsu.edu>. This guide includes discussion points that should be made explicit by the instructor and includes continual reference to the main ideas meant to be emphasized in the materials. The guide, along with faculty professional development, can hopefully allow teacher educators to better understand the intended curriculum and implement the materials

with high fidelity. Faculty professional development efforts have been established through free workshops held at professional conferences and week-long summer institutes. Evaluations of the week-long summer institutes in 2009 and 2010 suggest that the fifteen participants increased their confidence in their ability to engage prospective teachers in discussions about center and spread in a distribution, as well as randomness, sample size and variability.

Future Directions

For this study, we did not examine other sources of evidence of prospective teachers' development of understanding related to coordinating center and spread. Such data may include prospective teachers' responses to a variety of content and pedagogical questions posed throughout the chapters and perhaps pedagogical pre- and post-tasks such as interpreting students' work, designing tasks for students, creating a lesson plan. In fact, teacher educators at multiple institutions have collected sample work from prospective teachers on tasks from each of the

chapters. Analysis of this data with a focus on coordinating center and spread may yield additional findings that can help the field better understand of the development of prospective teachers' reasoning about center and spread.

Prospective teachers' familiarity with expected ranges of values, their propensity to use these ideas in conceptual statistical tasks, and their pedagogical implementation of coordination of center and spread are three different phenomena. As shown in this work and in other literature, the transference from the first of these to the latter two is problematic. Future versions of these materials may need to engage prospective teachers' further into the use of interval thinking about expectation and variation in a broader range of statistical tasks. More importantly, prospective teachers will need to be more consistently challenged to consider how to create tasks, pose questions, and facilitate classroom discussions aimed at engaging their own students in the coordination of center and spread.

Table 4
Sample Revisions in Chapter 1 to Better Facilitate Interval Reasoning

Text of Questions in Fall 2007	Text of Questions in Fall 2009
Q16. What do you notice about the distribution of average salaries? Where are the data clumped? What is the general spread of the data? How would you describe the shape?	Q20. Create a fully separated plot of the Average Teacher Salaries. Either stack the data vertically or horizontally. What do you notice about the distribution of average salaries? Where are the data clumped? What is the general spread of the data? How would you describe the shape?
Q17. Use the Divider tool and the Reference tool to highlight a clump of data that is "typical" and a particular value that seems to represent a "typical" salary. Justify why you highlighted a clump and identified a particular value as typical.	Q21. Use the Divider tool to shade the upper half of the data, which contains approximately 50% of the cases. Which states are in the upper half of the average salary range? What factors may contribute to the higher salaries in these states?
Q18. Drag the vertical divider lines to shade the upper half of the data, which contains approximately 50% of the cases. Which states are in the upper half of the average salary range? What factors may contribute to the higher salaries in these states?	Q22. Drag the vertical divider lines to shade the middle half of the data, which contains approximately 50% of the cases. Describe the spread of the data in the middle 50%. What might contribute to this spread?
	Q23. Drag the vertical divider lines to highlight a modal clump of data that is representative of a cluster that contains many data points. Explain why you chose that range as the modal clump.
	Q24. Use the Reference tool to highlight a particular value that seems to represent a "typical" salary. Justify why you identified a particular value as typical and how you may have used the range you identified as a modal clump to assist you.
	Q25. How can the use of the dividers to partition the data set into separate regions be useful for students in analyzing the spread, center, and shape of the distribution?

REFERENCES

- Batanero, C., Godino, J. D., & Roa, R. (2004). Training teachers to teach probability. *Journal of Statistics Education* 12(1). Retrieved from <http://www.amstat.org/publications/jse/v12n1/batanero.html>
- Canada, D. (2006). Elementary preservice teachers' conception of variation in a probability context. *Statistics Education Research Journal*, 5(1), 36-63. Retrieved from <http://www.stat.auckland.ac.nz/serj>
- Capraro, R. M. (2004). Statistical significance, effect size reporting, and confidence intervals: Best reporting strategies. *Journal for Research in Mathematics Education*, 35, 57-62.
- Clements, D. H. (2007). Curriculum research: Toward a framework for "research-based curricula". *Journal for Research in Mathematics Education*, 38, 35-70.
- delMas, R. C. (2004). A comparison of mathematical and statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 79-95). Dordrecht, The Netherlands: Kluwer.
- Fidler, F. (2006). Should psychology abandon *p*-values and teach CI's instead? Evidence-based reforms in statistics education. *Proceedings of International Conference on Teaching Statistics 7*. Retrieved from <http://www.stat.auckland.ac.nz/~iase>
- Franklin, C., Kader, G., Mewborn, D. S., Moreno, J., Peck, R., Perry, M., & Schaeffer, R. (2005). *A Curriculum Framework for K-12 Statistics Education: GAISE report*. Alexandria, VA: American Statistical Association. Retrieved from <http://www.amstat.org/education/gaise>
- Friel, S., O'Connor, W., & Mamer, J. (2006). More than "medianmode" and a bar graph: What's needed to have a statistics conversation? In G. Burrill (Ed.), *Thinking and reasoning with data and chance: Sixty-eighth yearbook* (pp. 117-137). Reston, VA: National Council of Teachers of Mathematics.
- Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than *p* values: Estimation rather than hypothesis testing. *British Medical Journal (Clinical Research Ed)*, 292, 746-750.
- Garfield, J. B. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education*, 10(3). Retrieved from <http://www.amstat.org/publications/jse/v10n3/garfield.html>
- Garfield, J. B. (2003). Assessing statistical reasoning. *Statistics Education Research Journal*, 2(1), 22-38. Retrieved from [http://www.stat.auckland.ac.nz/~iase/serj/SERJ2\(1\).pdf](http://www.stat.auckland.ac.nz/~iase/serj/SERJ2(1).pdf)
- Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66, 64-74.
- International Committee of Medical Journal Editors. (1997). Uniform requirements for manuscripts submitted to biomedical journals. *Annals of Internal Medicine* 126 (1), 36-47, [online] Available at <http://www.annals.org/cgi/content/full/126/1/36>.
- Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, 33, 259-289.
- Leavy, A. M. (2010). The challenge of preparing preservice teachers to teach informal inferential reasoning. *Statistics Education Research Journal*, 9(1), 46-67, Retrieved from <http://www.stat.auckland.ac.nz/serj>.
- Lee, H. S., & Hollebrands, K. F. (2008a). Preparing to teach mathematics with technology: An integrated approach to developing technological pedagogical content knowledge. *Contemporary Issues in Technology and Teacher Education*, 8(4). Retrieved from <http://www.citejournal.org/vol8/iss4/mathematics/article1.cfm>
- Lee, H. S., & Hollebrands, K. F. (2008b). Preparing to teach data analysis and probability with technology. In C. Batanero, G. Burrill, C. Reading, & A. Rossman (Eds.), *Joint ICMI/IASE Study: Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education. Proceedings of the ICMI Study 18 and 2008 IASE Round Table Conference*. Retrieved from http://www.ugr.es/~icmi/iase_study/Files/Topic3/T3P4_Lee.pdf
- Lee, H. S., Hollebrands, K. F., & Wilson, P. H. (2010). *Preparing to teach mathematics with technology: An integrated approach to data analysis and probability*. Dubuque, IA: Kendall Hunt.
- Lee, H. S., & Mojica, G. F. (2008). Examining teachers' practices: In what ways are probabilistic reasoning and statistical investigations supported? . In C. Batanero, G. Burrill, C. Reading, & A. Rossman (Eds.), *Joint ICMI/IASE study: Teaching statistics in school mathematics. Challenges for teaching and teacher education. Proceedings of the ICMI Study 18 and 2008 IASE Round Table Conference*. Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/rt08/T2P14_Lee.pdf.
- Lesser, L. M., & Groth, R. (2008). Technological pedagogical content knowledge in statistics. In *Electronic Proceedings of the Twentieth Annual International Conference on Technology in Collegiate Mathematics* (pp. 148-152). Retrieved from <http://archives.math.utk.edu/ICTCM/VOL20/S118/paper.pdf>
- Makar, K., & Confrey, J. (2005). "Variation-talk": Articulating meaning in statistics. *Statistics Education Research Journal*, 4(1), 27-54. Retrieved from <http://www.stat.auckland.ac.nz/serj>
- National Council of Teachers of Mathematics. (2000). *Principles and Standards for School Mathematics*. Reston, VA: Author.
- Powell, A. B., Francisco, J. M., & Maher, C. A. (2003). An analytical model for studying the development of learners' mathematical ideas and reasoning using videotape data. *Journal of Mathematical Behavior*, 22, 405-435.
- Reading, C., & Shaughnessy, J. M. (2000). Student perceptions of variation in a sampling situation. In T. Nakahara & M. Koyama (Eds.), *Proceedings of the 24th annual conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 89-96). Hiroshima, Japan: Hiroshima University.
- Reading, C., & Shaughnessy, J. M. (2004). Reasoning about variation. In J. Garfield & D. Ben-Zvi (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 201-226). Dordrecht, The Netherlands: Kluwer.
- Saldanha, L., & Thompson, P. W. (2002). Conceptions of sample and their relationships to statistical inference. *Educational Studies in Mathematics*, 51, 257-270.
- Shaughnessy, J. M. (2006). Research on students' understanding of some big concepts in statistics. In G. Burrill (Ed.), *Thinking*

and reasoning with data and chance: Sixty-eighth yearbook (pp. 77-98). Reston, VA: National Council of Teachers of Mathematics.

Stohl, H. (2005). Probability in teacher education and development. In G. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 345-366). Dordrecht, The Netherlands: Kluwer.

Watson, J. M., Callingham, R. A., & Kelly, B. A. (2007). Students' appreciation of expectation and variation as a foundation for statistical understanding. *Mathematical Thinking and Learning*, 9, 83-130.